Statistics

Roya Najafi Vosough
Ph.D. in Biostatistics

What is in this workshop

- Basics of Statistics
- How to use SPSS
- How to get descriptive statistics
- How to get inferential statistics

Statistics

- The science of collection, organization, summarize, analysis, interpretation of data.

Two areas of statistics:

- Descriptive statistics

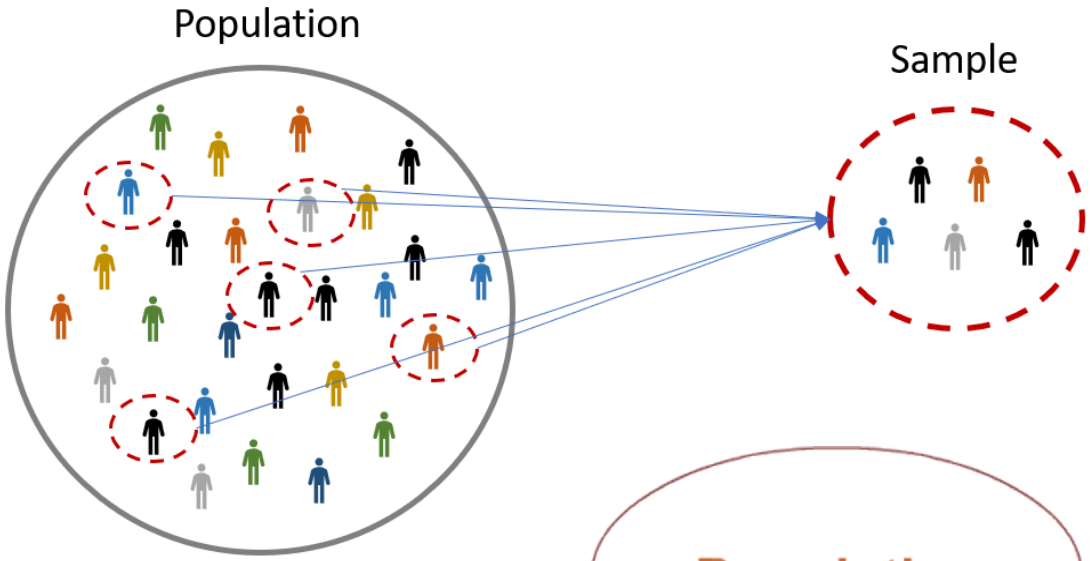
  (which summarize some characteristic of a sample)

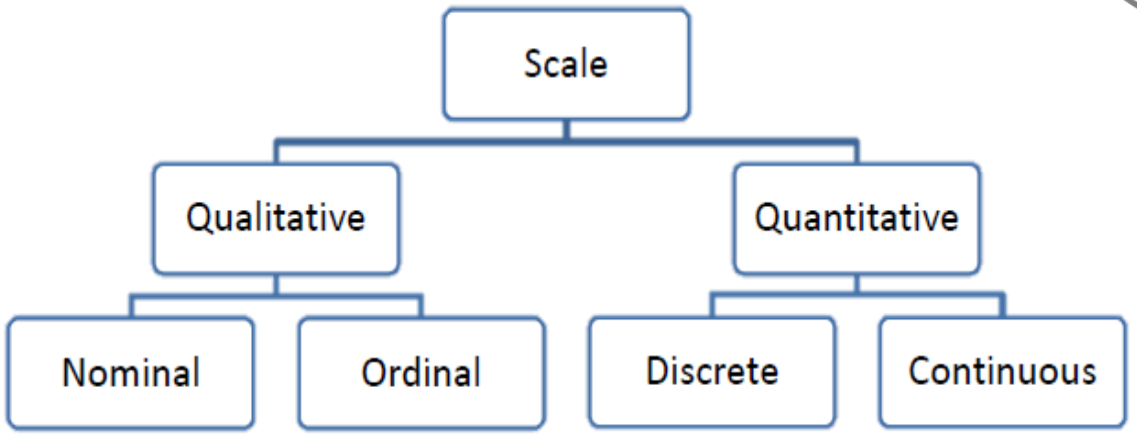
- Inferential statistics

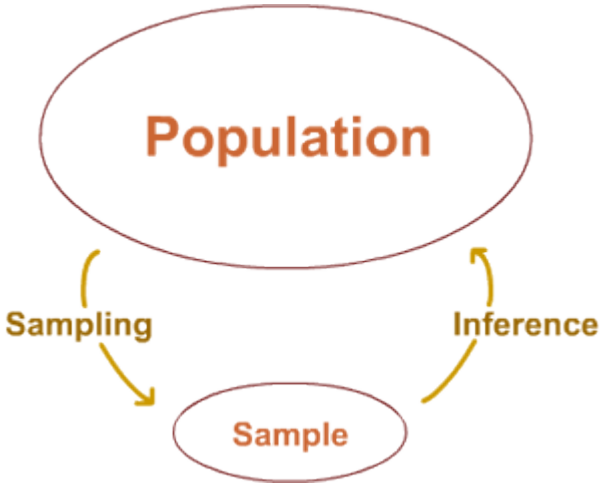
  (which test for significant differences between groups and/or significant relationships among variables within the sample)

Basic Terms

- Population
- Sample
- Variables



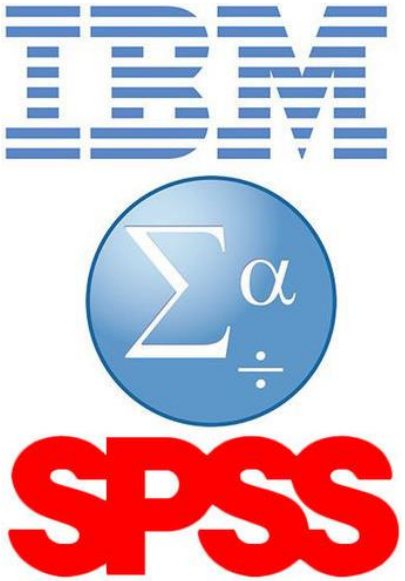




Data Analytics Tools

What is SPSS ?

- SPSS stands for Statistical Package for the Social Sciences
- SPSS was made to be easier to use then other statistical software like S-Plus, R, or SAS.
- The newest version of SPSS is SPSS 26.0.



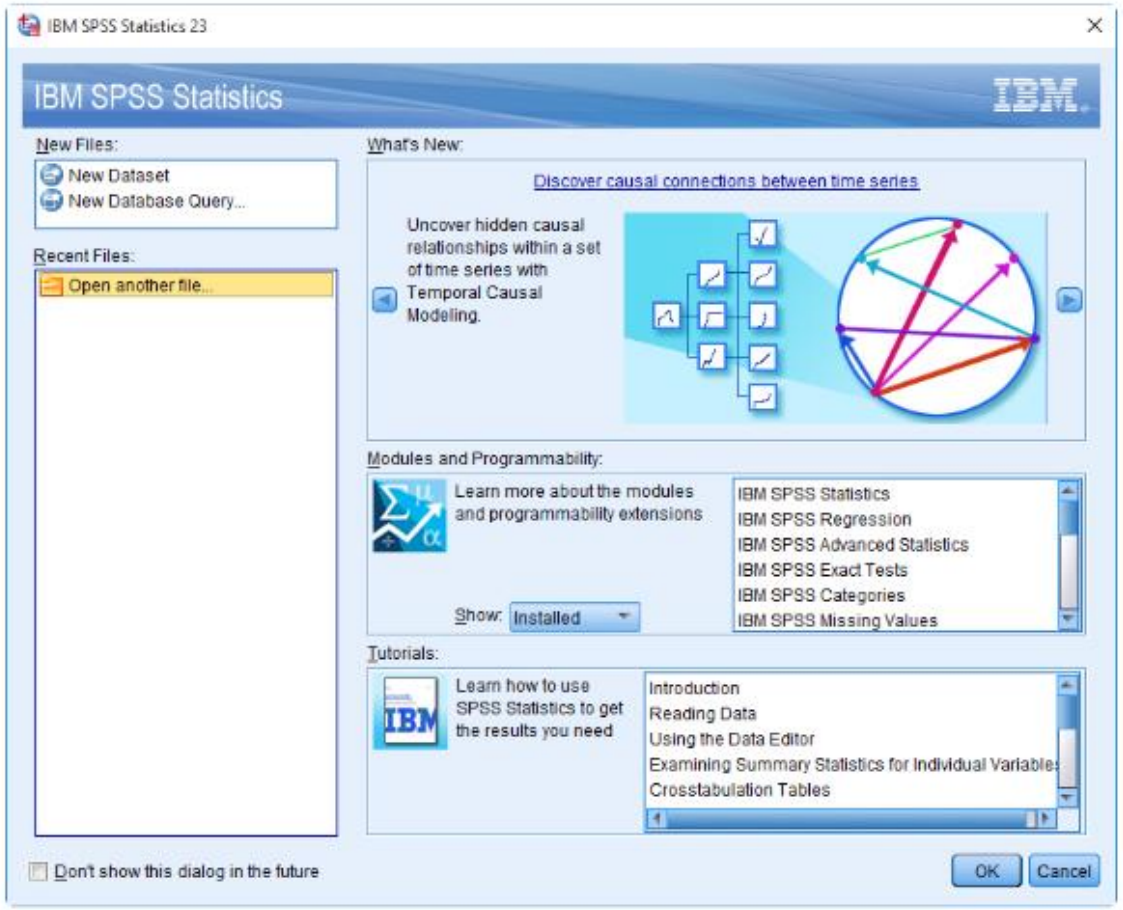
Uses for SPSS

- Data management
  - Defining variables
  - Coding values
  - Entering and editing data
  - Creating new variables
  - Recoding variables
  - Selecting cases
- Data analysis

How to open SPSS

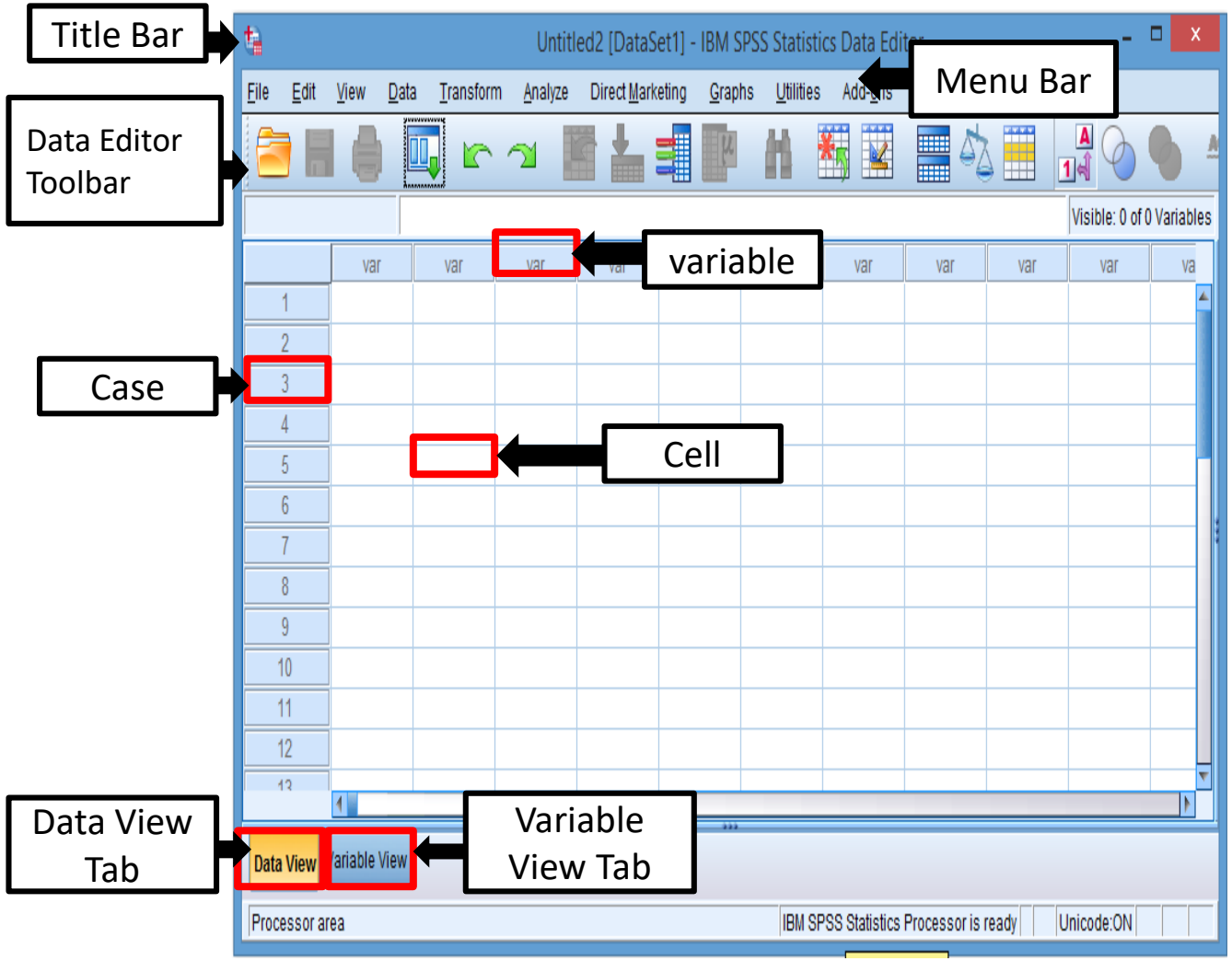
Double click the SPSS icon on the desktop

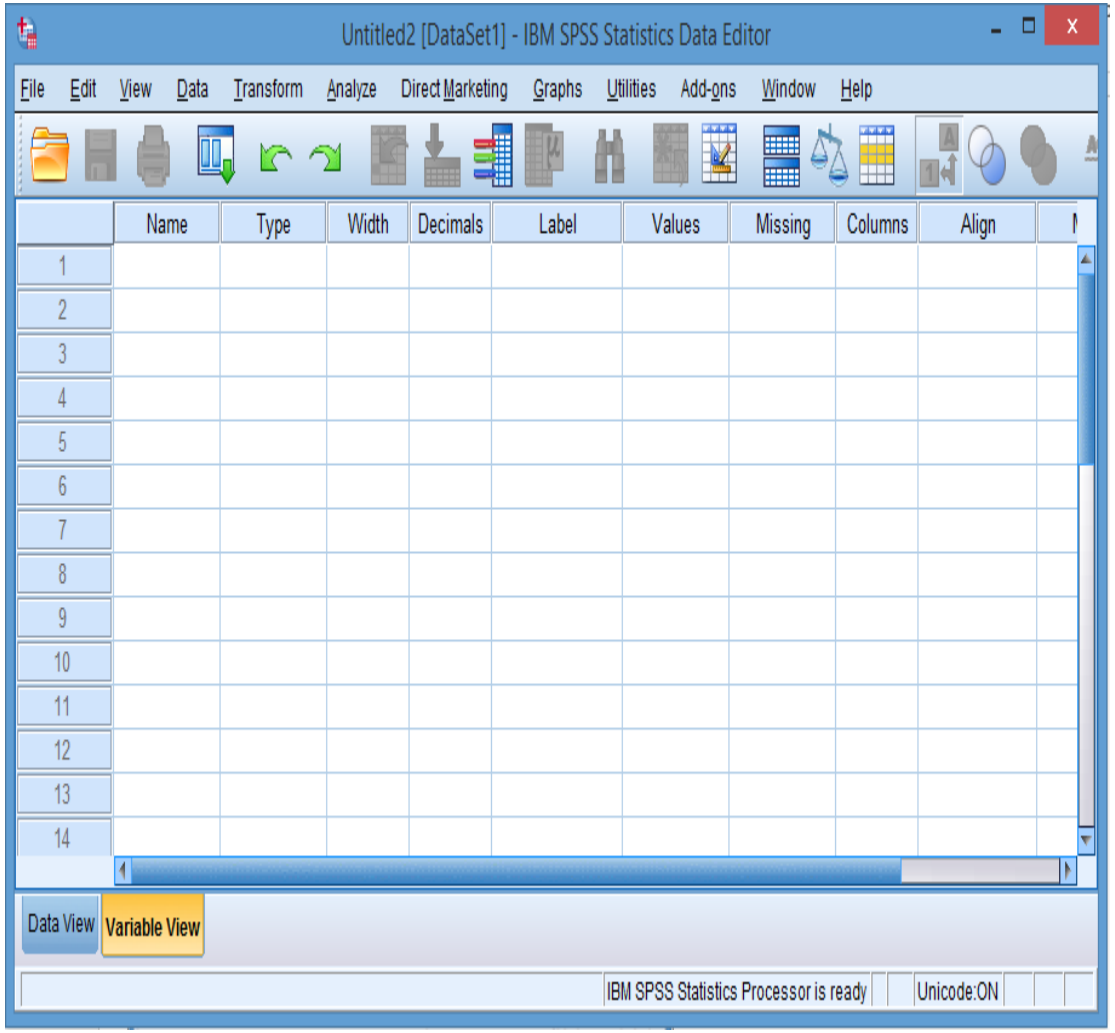
Start Menu → Programs → SPSS for windows → SPSS



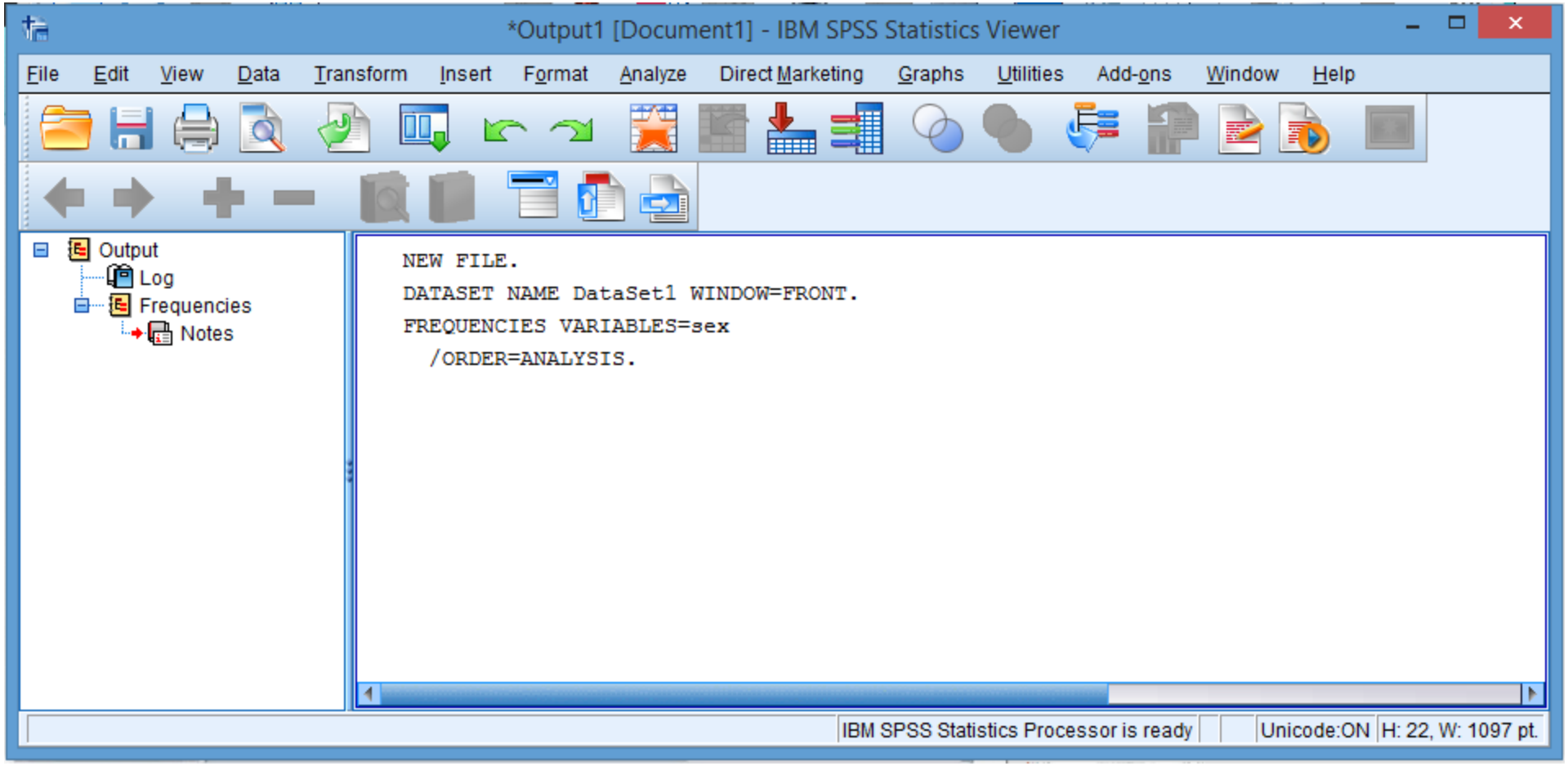
The two main windows

Data editor


Title Bar
Data Editor Toolbar
Case
Data View Tab
Untitled2 [DataSet1] - IBM SPSS Statistics Data Editor
Menu Bar
variable
Cell
Variable View Tab
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons
Visible: 0 of 0 Variables
var
Data View
Variable View
Processor area
IBM SPSS Statistics Processor is ready
Unicode:ON


Untitled2 [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Name
Type
Width
Decimals
Label
Values
Missing
Columns
Align
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

Output viewer



Descriptive statistics

- Collection
- Organization
- Summarize data
  - Tables
  - Graphs
  - Measures of Central Tendency
  - Index of dispersion

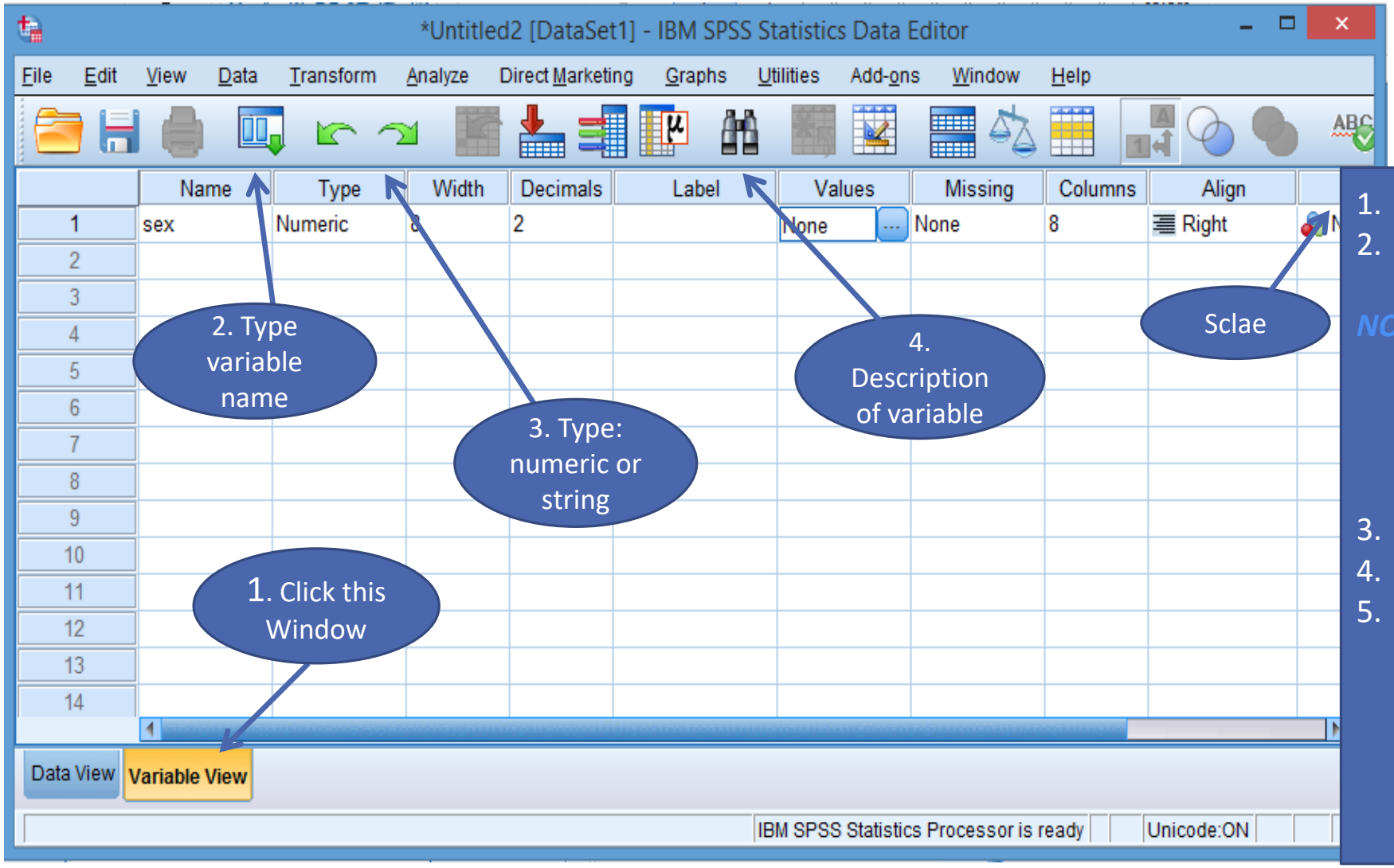
Example

PATIENTS FORM

ID:

DEMOGRAPHIC INFORMATION:		
Sex:	1) Male ☐	2) Female ☐
Age (year):	☐☐	
Diabetes History	1) yes ☐	2) No ☐
Hypertension history	1) yes ☐	2) No ☐
Hyperlipidemia history	1) yes ☐	2) No ☐
Smoking history	1) yes ☐	2) No ☐

Enter variables

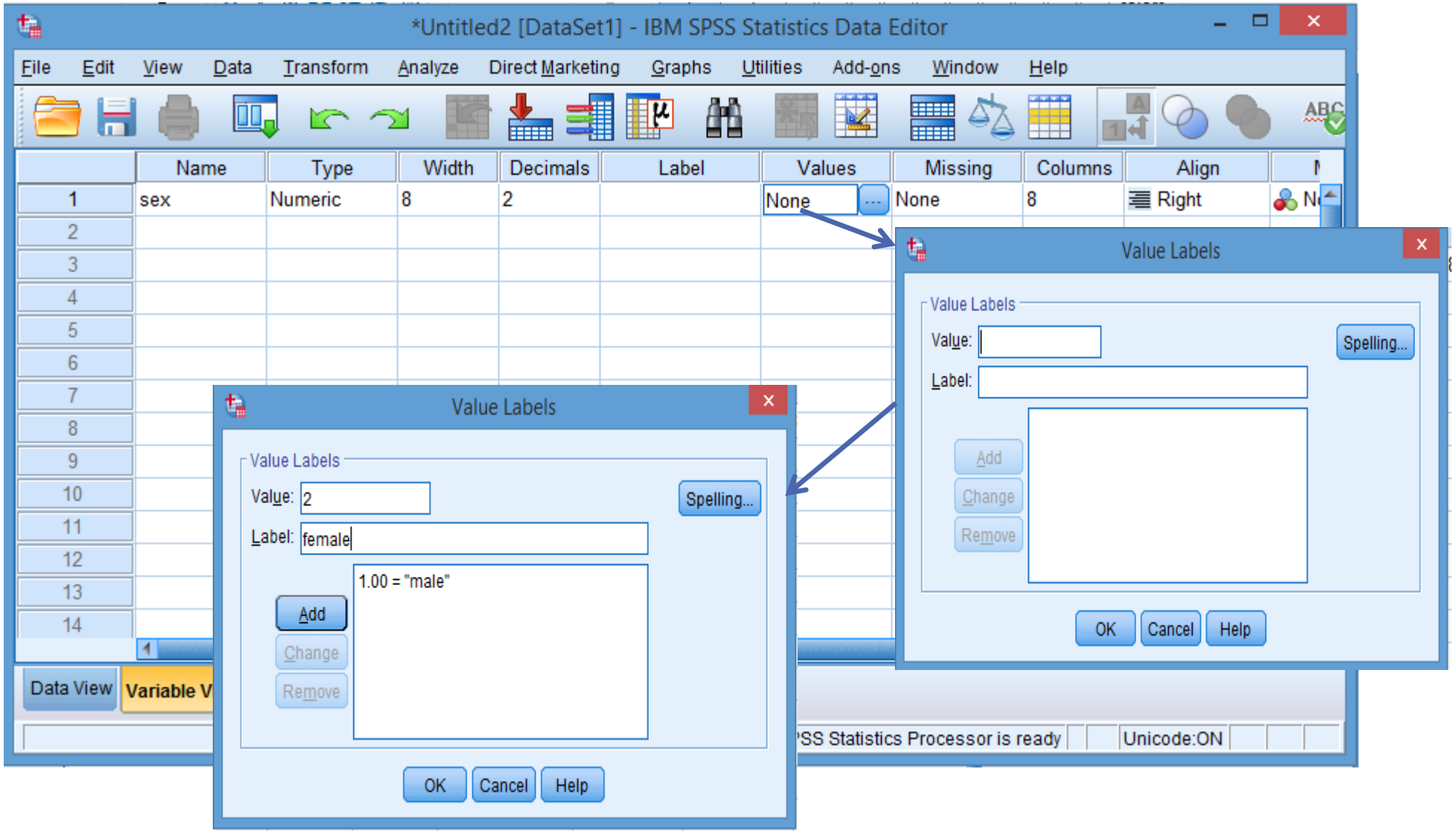


1. Click Variable View
2. Type variable name under Name column (e.g. Sex).

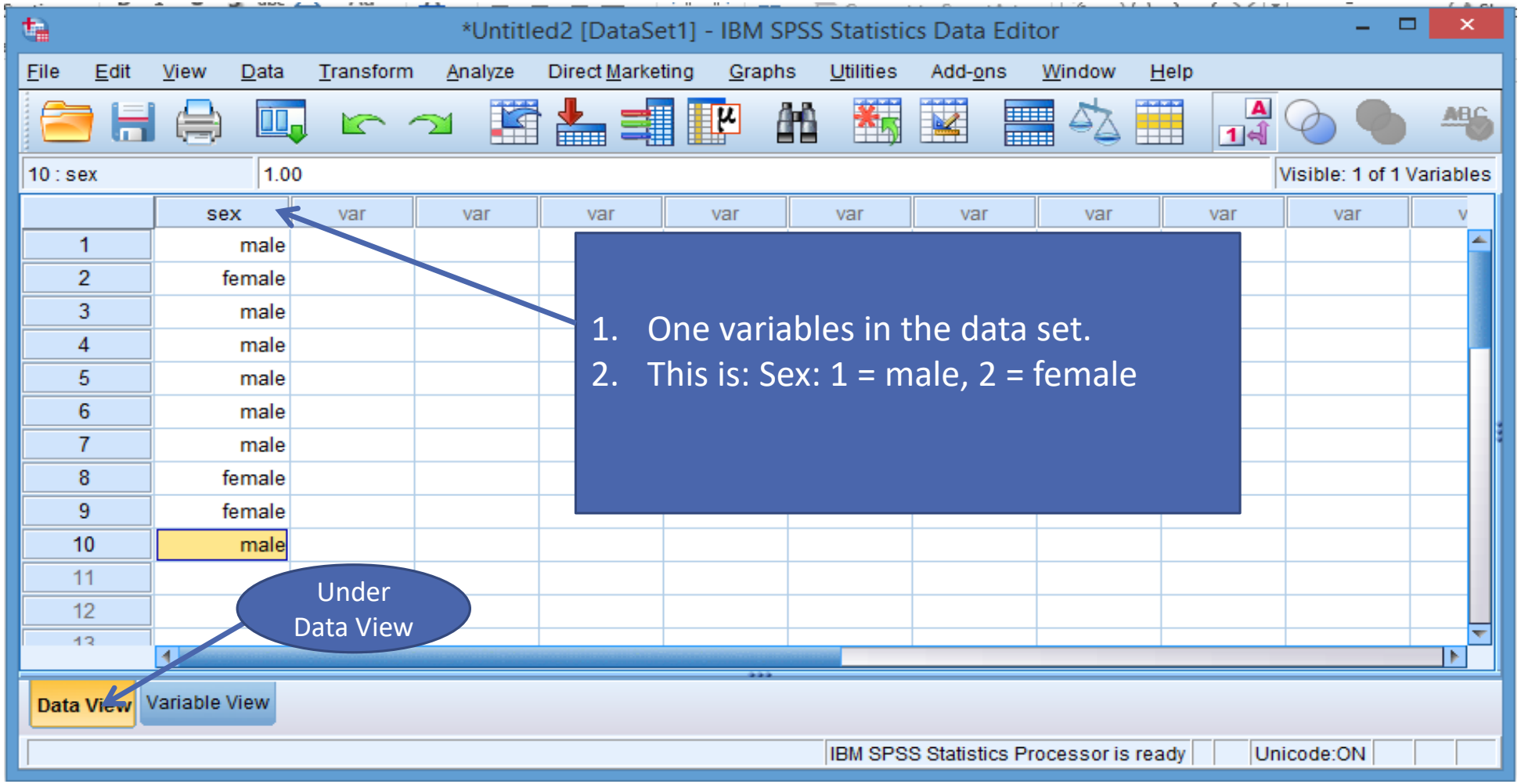
NOTE: Variable name can be 64 bytes long, and the first character must be a letter or one of the characters @, #, or $.

3. Type: Numeric, string, etc.
4. Label: description of variables.
5. Measure: Nominal, Ordinal, Scale

Enter variables



Enter cases


*Untitled2 [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
10 : sex
1.00
Visible: 1 of 1 Variables
sex
male
female
male
male
male
male
male
female
female
male
1. One variables in the data set.
2. This is: Sex: 1 = male, 2 = female
Under Data View
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

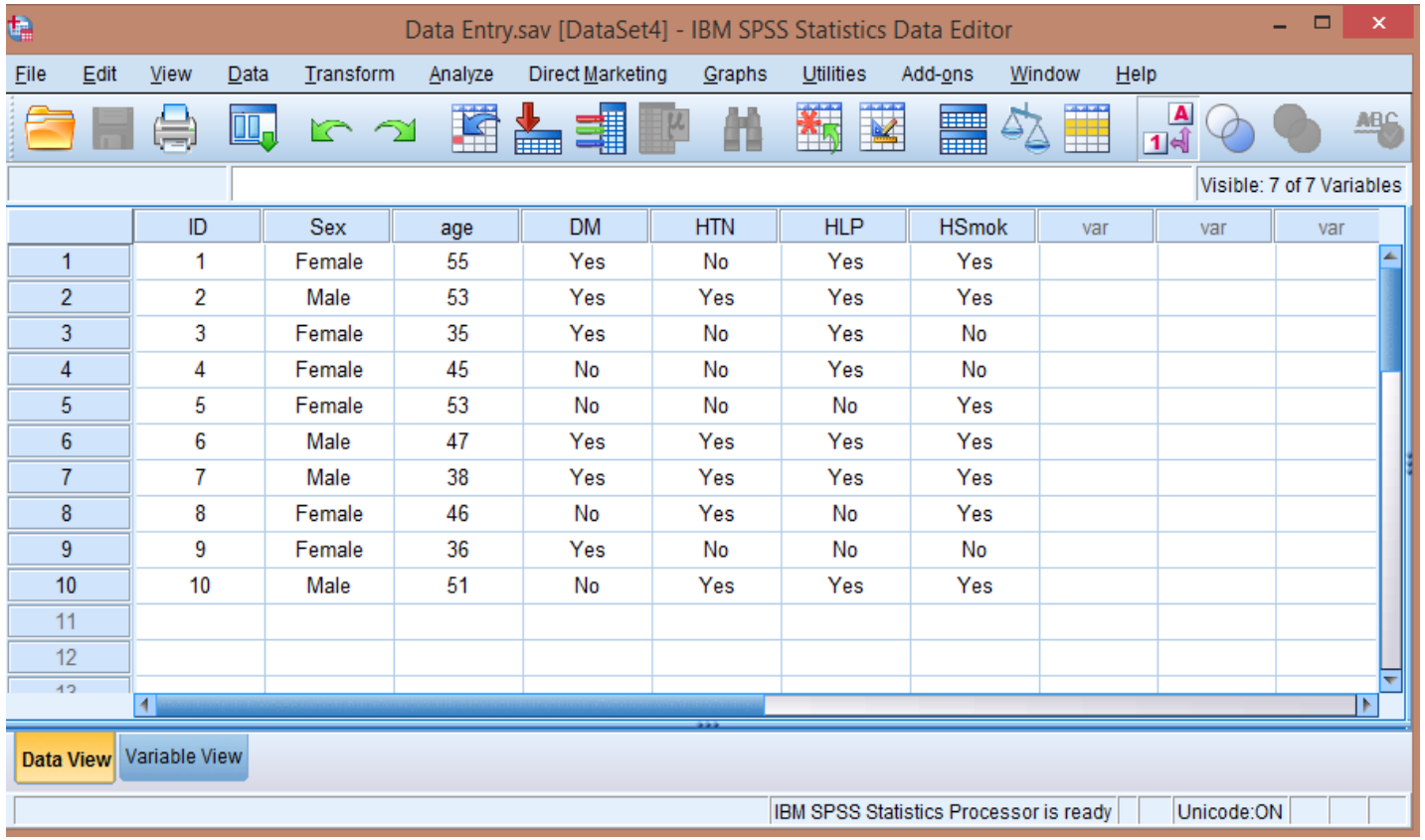
How would you put the following information in to SPSS ??

Sex	Age	Diabetes History	Hypertension history	Hyperlipidemia history	Smoking history
Female	55	Yes	No	Yes	Yes
Male	53	Yes	Yes	Yes	Yes
Female	35	Yes	No	Yes	No
Female	45	No	No	Yes	No
Female	53	No	No	No	Yes
Male	47	Yes	Yes	Yes	Yes
Male	38	Yes	Yes	Yes	Yes
Female	46	No	Yes	No	Yes
Female	36	Yes	No	No	No
Male	51	No	Yes	Yes	Yes

Data used in the workshop

Variable:

- Sex
- Age
- Diabetes History
- Hypertension history
- Hyperlipidemia history
- Smoking history

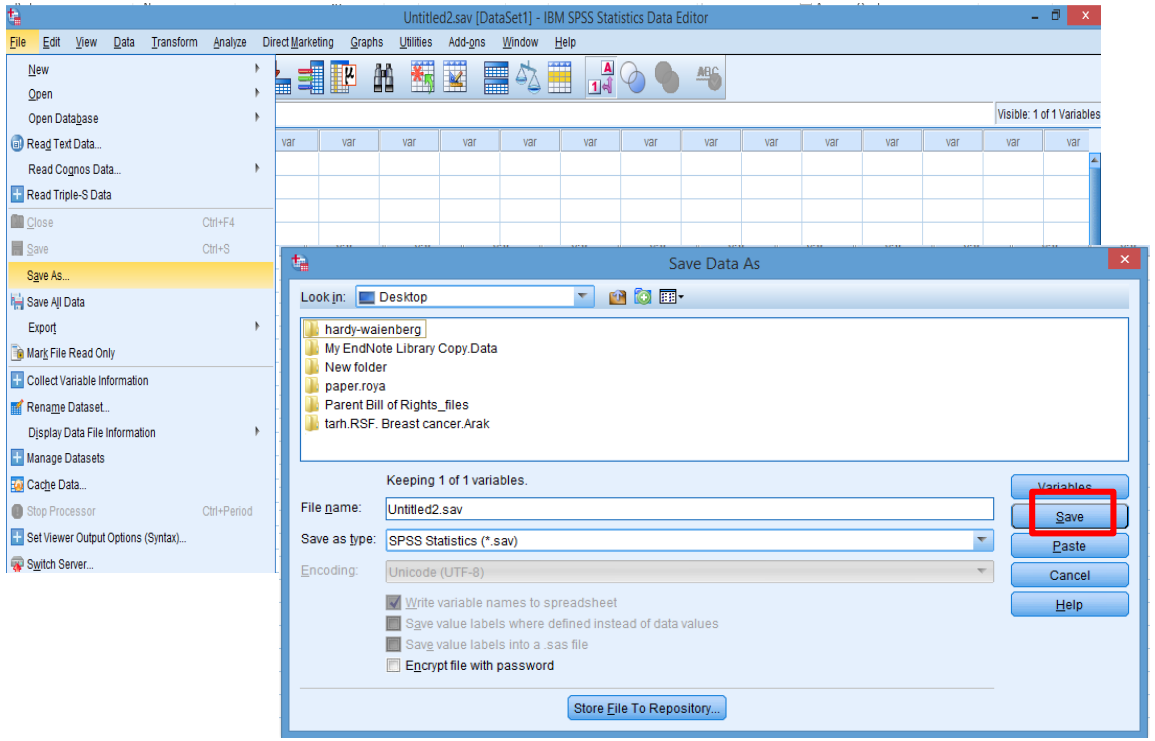


	ID	Sex	age	DM	HTN	HLP	HSmok	var	var	var
1	1	Female	55	Yes	No	Yes	Yes			
2	2	Male	53	Yes	Yes	Yes	Yes			
3	3	Female	35	Yes	No	Yes	No			
4	4	Female	45	No	No	Yes	No			
5	5	Female	53	No	No	No	Yes			
6	6	Male	47	Yes	Yes	Yes	Yes			
7	7	Male	38	Yes	Yes	Yes	Yes			
8	8	Female	46	No	Yes	No	Yes			
9	9	Female	36	Yes	No	No	No			
10	10	Male	51	No	Yes	Yes	Yes			
11										
12										

Saving the file

The file must always be saved in order to save the work that has been done to date:

- File/Save as
- Move to the target directory
- Enter a file name
- Save

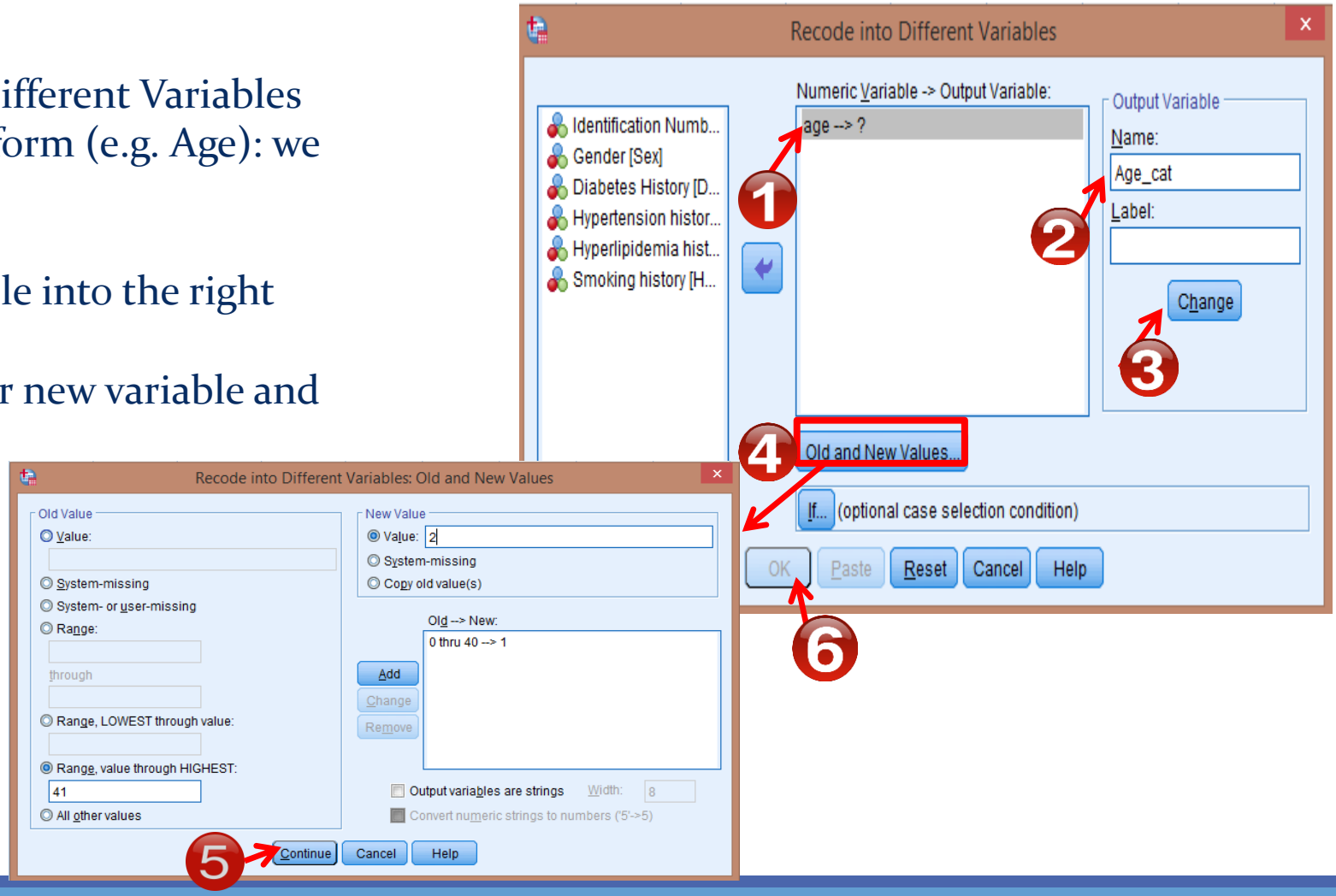


Descriptive statistics

- Collection
- Organization
- Summarize data
  - Tables
  - Graphs
  - Measures of Central Tendency
  - Index of dispersion

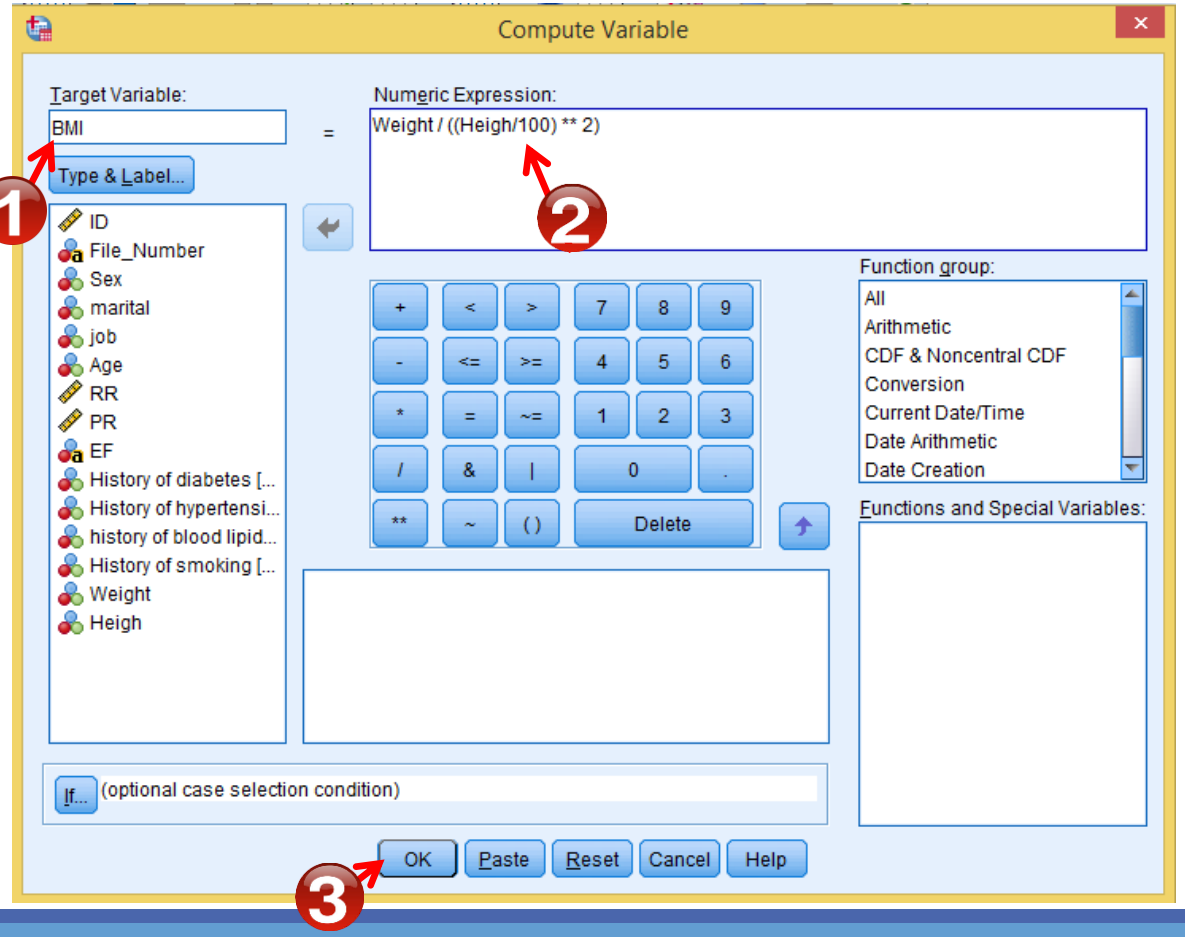
Variable transformation

1. Select Transform ⟶ Recode into Different Variables
2. Select variable that you want to transform (e.g. Age): we want
   1: =<40 and 2 :>41
3. Click Arrow button to put your variable into the right window
4. Under Output Variable: type name for new variable and label, then click Change
5. Click Old and New Values



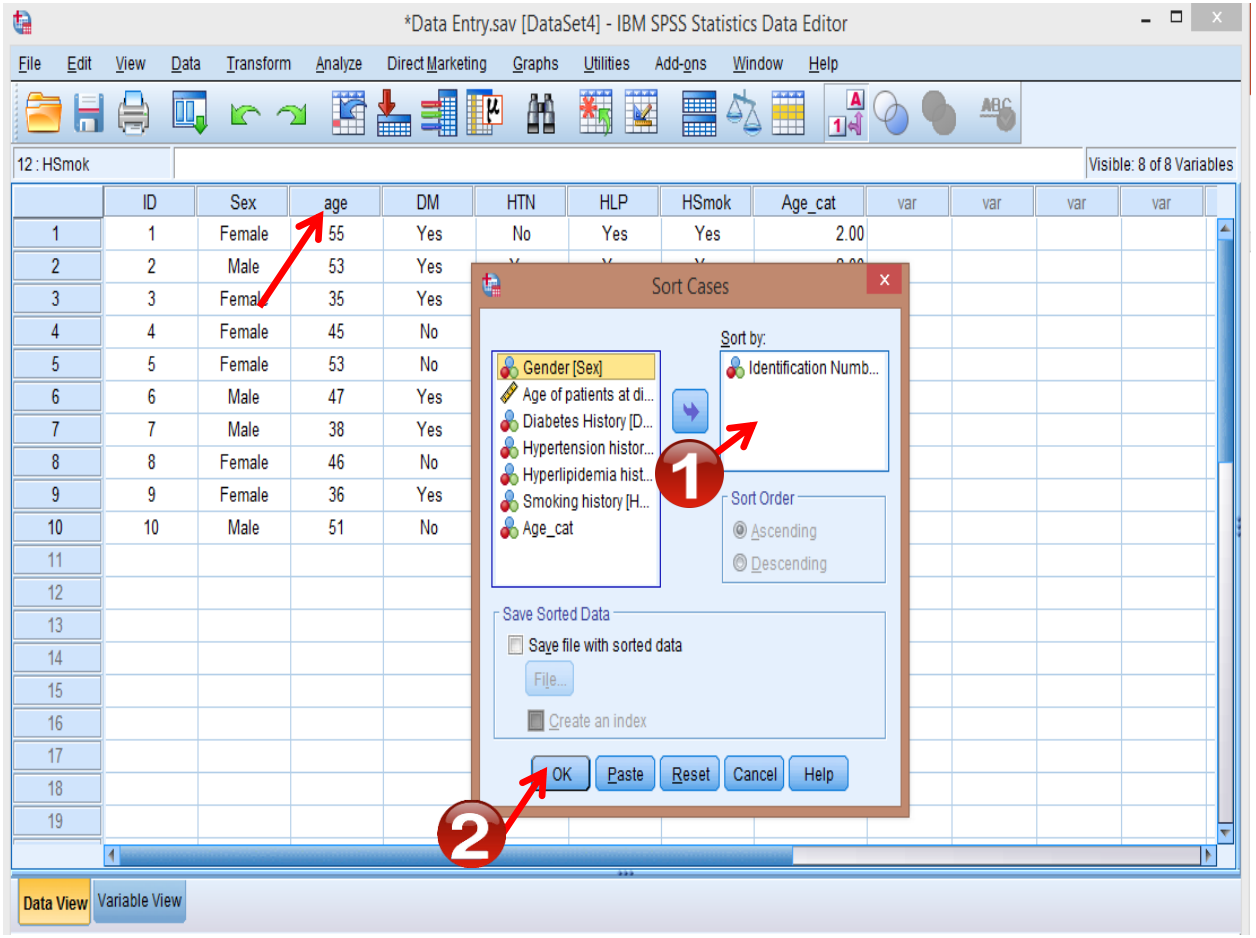
Variable transformation

- Compute variable
- Example 1. Create a new variable: BMI
- Use Weight, Height
- Go to Transform ⟶ Compute Variable



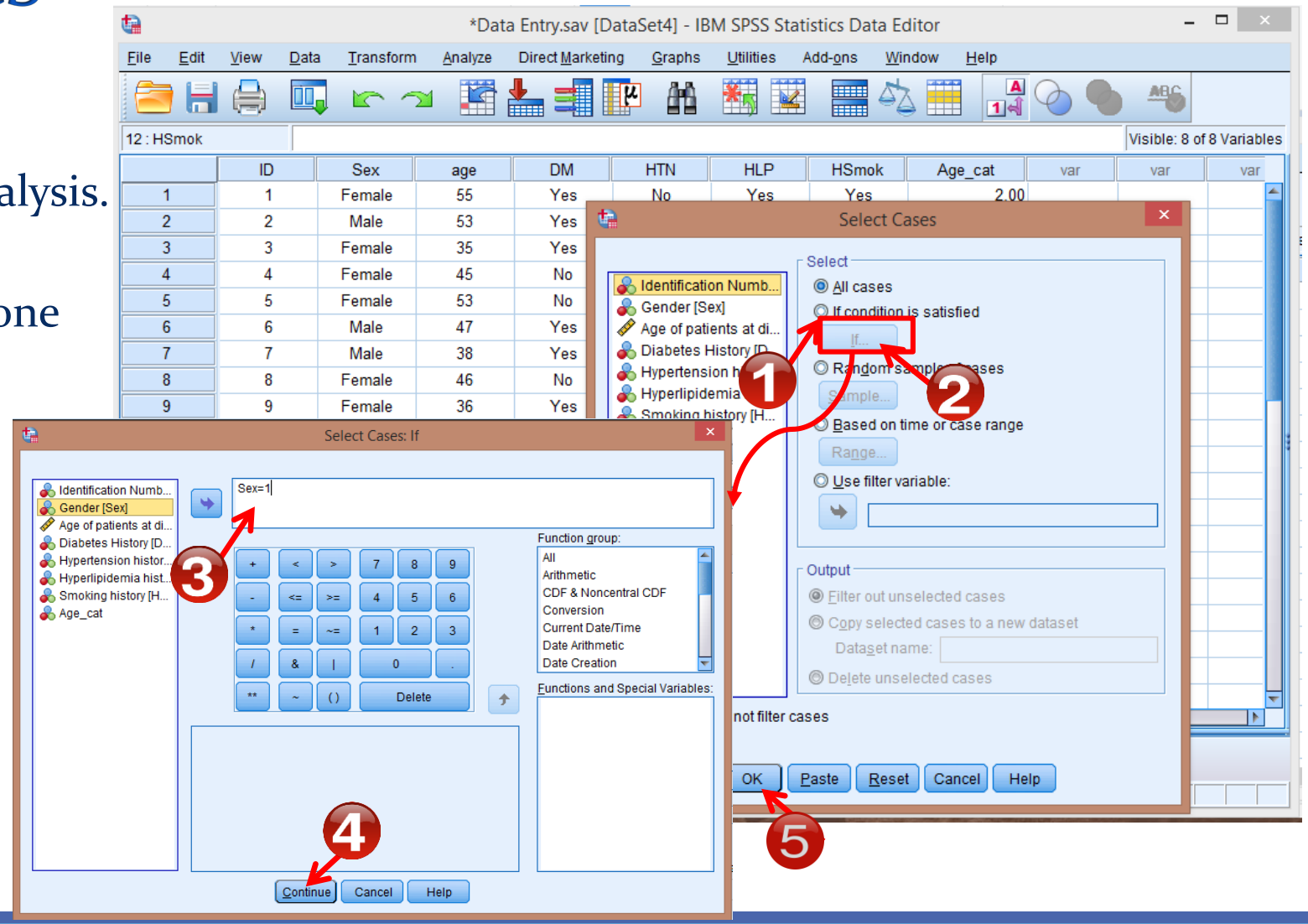
Sort and select cases

- Sort cases
  - Sort cases by variables: Data ⟶ Sort Cases
  - You can use Sort Cases to find missing.



Sort and select cases

- Select cases
  - Example 1. Select Females for analysis.
  - Go to Data ⟶ Select Cases
  - Under Select: check the second one
  - Click If button

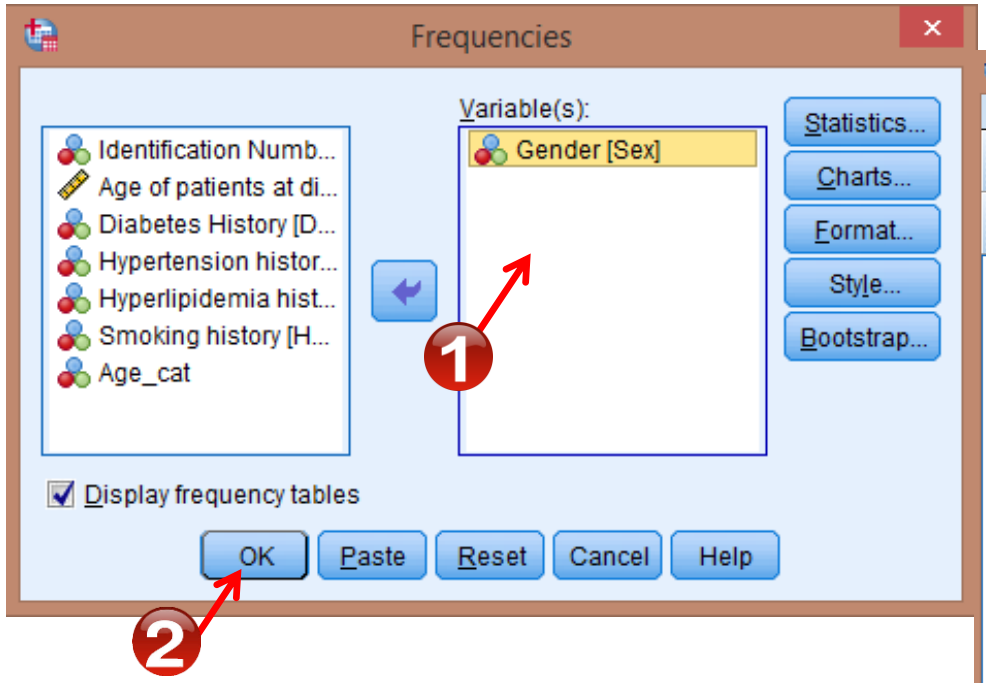


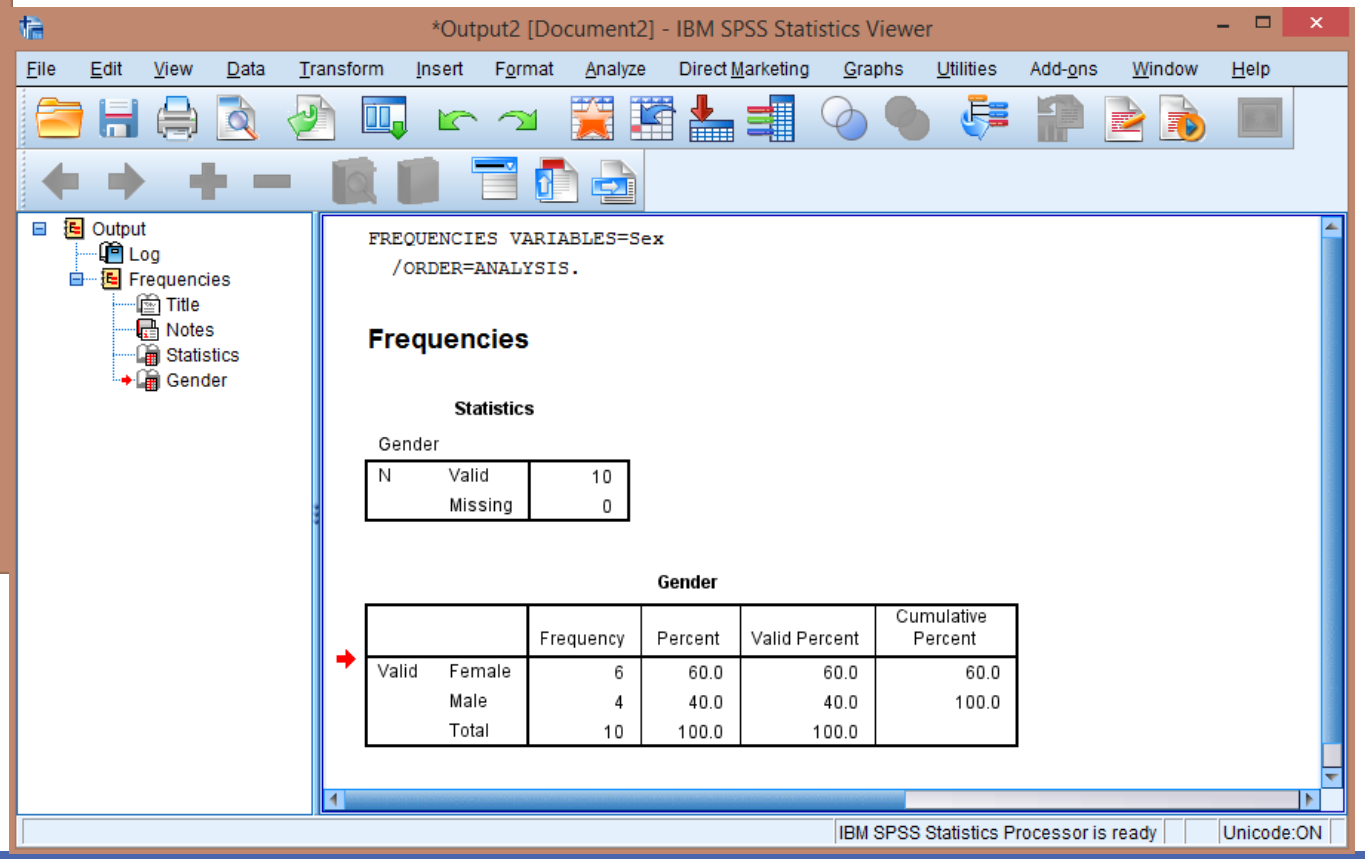
Descriptive statistics

- Collection
- Organization
- Summarize data
  - Tables
  - Graphs
  - Measures of Central Tendency
  - Index of dispersion

Descriptive statistics: Tables

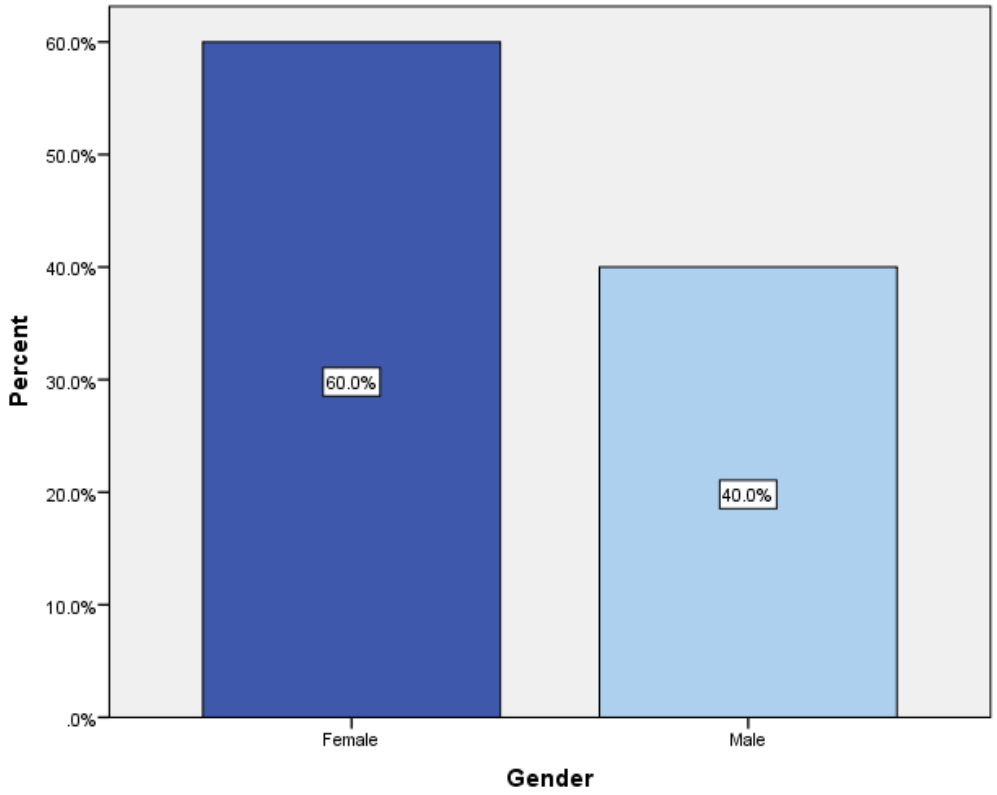
Analyze → Descriptive statistics → Frequency

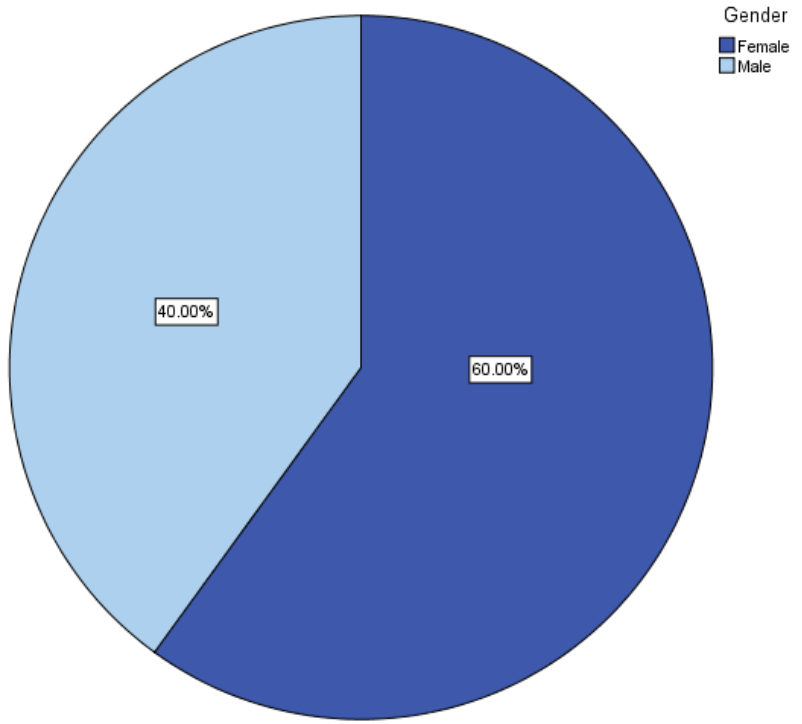




Descriptive statistics: Graphs

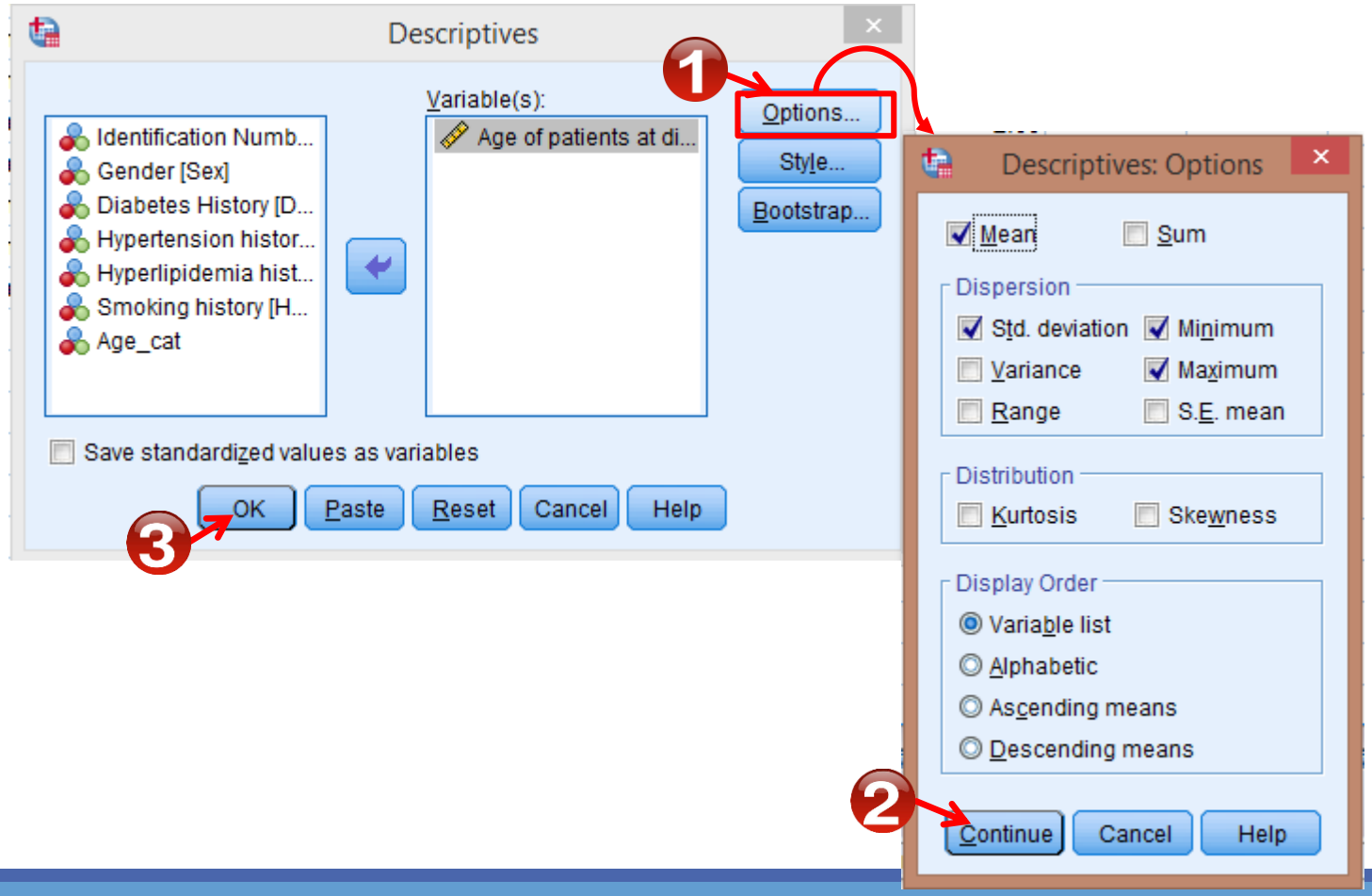
Graphs → Legacy Dialogs → Bar, Pie, …





Descriptive statistics: Measures of Central Tendency & Index of dispersion

Analyze → Descriptive statistics → Descriptive



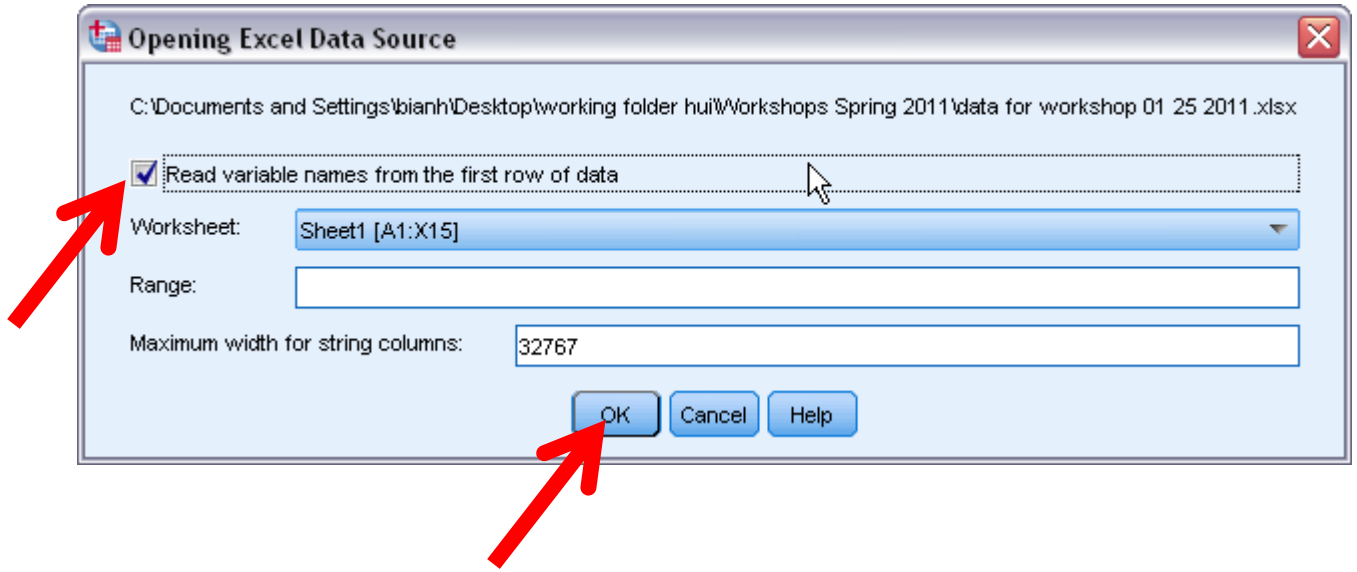
Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Age of patients at diagnosis	10	35	55	45.90	7.385
Valid N (listwise)	10				

Import data from Excel

- Select File ⟶ Open ⟶ Data
- Choose **Excel** as file type
- Select the file you want to import
- Then click Open



Inferential statistics

- Estimating population parameter from sample data
- Hypothesis Testing
- Statistical Methods of Analysis
  - Parametric
  - Non-Parametric

Chi-Square Statistic (Pearson's chi-square test)

- We perform this test when we want to determines whether there is an association between categorical variables.

- Assumptions
  1. Independent observations
  2. For a 2 by 2 table, all expected frequencies > 5.
     For a larger table, no more than 20% of all cells may have an expected frequency < 5 and all expected frequencies > 1.

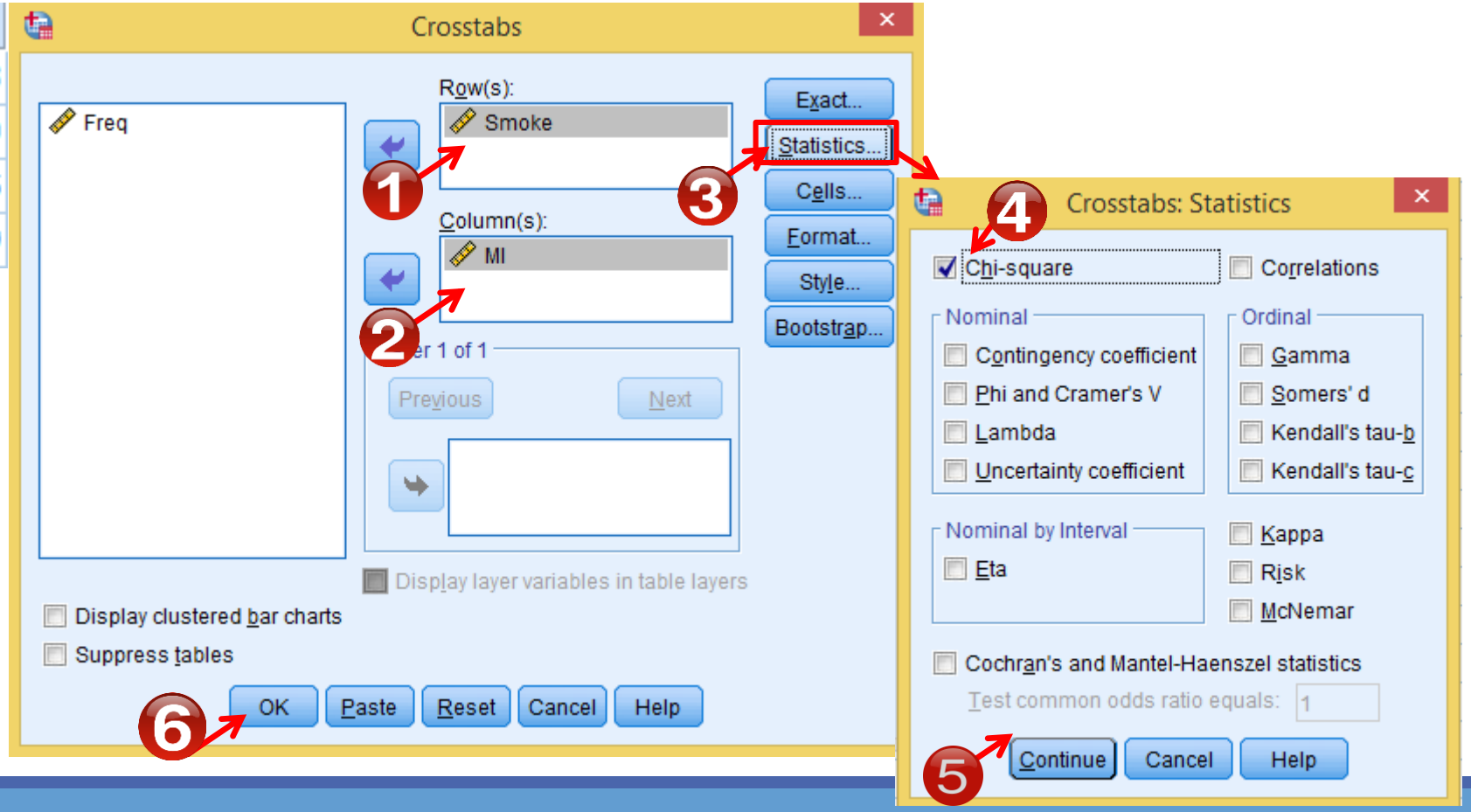
Null hypothesis: (Variable 1) is independent of (Variable 2).
(Variable 1) is not associated with (Variable 2).

If the p-value is less than alpha, you reject the null hypothesis.

Chi-Square Statistic (Pearson's chi-square test)

Analyze → Descriptive Statistics → Crosstabs

Smoke	MI	Freq
1	1	1413
1	2	1250
2	1	56
2	2	219



Chi-Square Statistic (Pearson's chi-square test)

Smoke * MI Crosstabulation

Count

		MI		
		Yes	No	Total
Smoke	Yes	1413	1250	2663
	No	56	219	275
Total		1469	1469	2938

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	106.592[a]	1	.000		
Continuity Correction[b]	105.288	1	.000		
Likelihood Ratio	113.243	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	106.555	1	.000		
N of Valid Cases	2938				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 137.50.

b. Computed only for a 2x2 table

Independent Sample T-Test (Independent T-Test or Two-Sample T-Test)

- We perform this test when we want to compare the mean of two different samples.
- Assumptions
  1. Independent observations
  2. Normality (Kolmogorov–Smirnov test/ Shapiro–Wilk test)
  3. Homogeneity of variance (Levene's test)

Null hypothesis: The two population means are equal.
The difference between the two population means is equal to 0.

If the p-value is less than alpha, you reject the null hypothesis.

Normality test

- Kolmogorov–Smirnov test

Analyze → Nonparametric Tests → Legacy Dialogs → 1-Sample K-S

- Shapiro–Wilk test

Analyze → Descriptive Statistics → Explore → Plots → Normality plots with tests

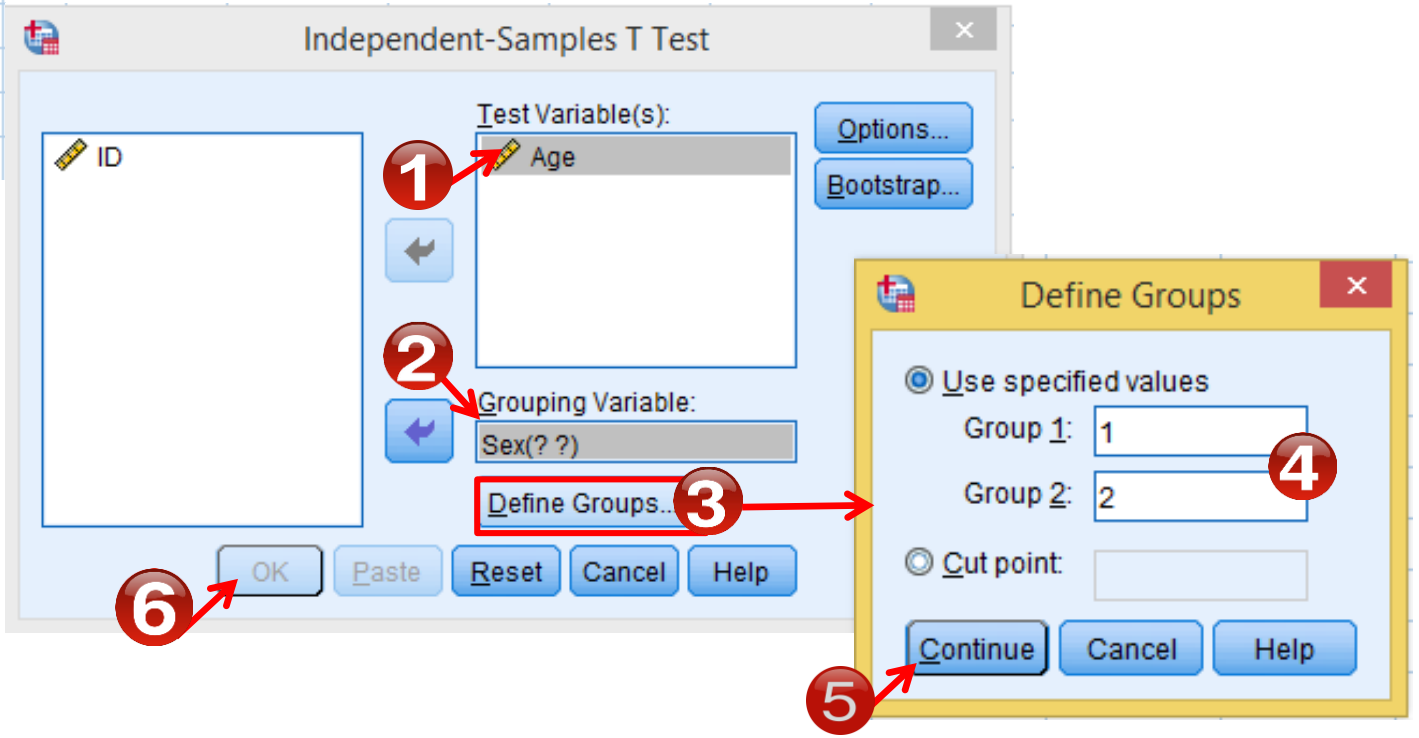
Null hypothesis: The data is normally distributed

If the p-value is less than alpha, you reject the null hypothesis.

Independent Sample T-Test (Independent T-Test or Two-Sample T-Test)

Analyze ⟶ Compare Means ⟶ Independent-Samples T-Test

ID	Sex	Age
1	male	44.00
2	male	80.00
3	male	79.00
4	male	70.00
5	male	73.00
⋮		
35	female	79.00
36	female	80.00
37	female	50.00
38	female	76.00
39	female	64.00
40	female	64.00



Independent Sample T-Test (Independent T-Test or Two-Sample T-Test)

Assumptions (Normality)

Sex = male

One-Sample Kolmogorov-Smirnov Test[a]

		Age
N		20
Normal Parameters[b,c]	Mean	68.6000
	Std. Deviation	13.02791
Most Extreme Differences	Absolute	.103
	Positive	.052
	Negative	-.103
Test Statistic		.103
Asymp. Sig. (2-tailed)		.200[d,e]

Sex = female

One-Sample Kolmogorov-Smirnov Test[a]

		Age
N		20
Normal Parameters[b,c]	Mean	71.9000
	Std. Deviation	10.48257
Most Extreme Differences	Absolute	.130
	Positive	.130
	Negative	-.102
Test Statistic		.130
Asymp. Sig. (2-tailed)		.200[d,e]

T-Test

Group Statistics

	Sex	N	Mean	Std. Deviation	Std. Error Mean
Age	male	20	68.6000	13.02791	2.91313
	female	20	71.9000	10.48257	2.34397

Assumptions (Homogeneity of variance)

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Age	Equal variances assumed	1.093	.302	-.883	38	.383	-3.30000	3.73905	-10.86932	4.26932
	Equal variances not assumed			-.883	36.336	.383	-3.30000	3.73905	-10.88072	4.28072

Example

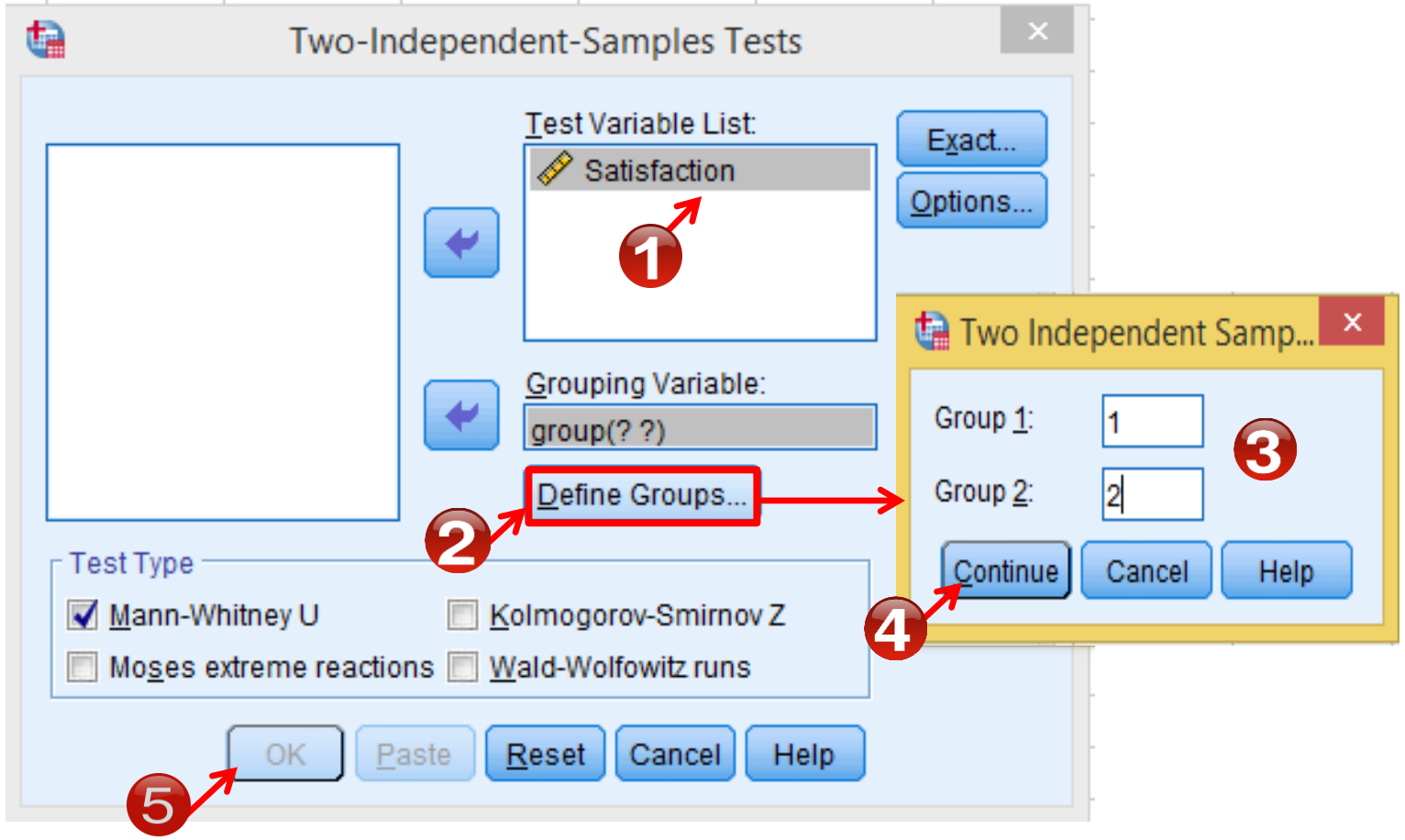
- The Effect of Nursing Consultation on Satisfaction of Patient's Families at the Cardiac Surgical Intensive Care Unit.

group	Satisfaction	group	Satisfaction
Control	2.29	Case	4.33
Control	2.43	Case	4.05
Control	2.52	Case	4.19
Control	2.52	Case	4.14
Control	2.33	Case	4.43
Control	2.62	Case	4.38
Control	2.33	Case	4.57
Control	2.48	Case	4.38
Control	2.67	Case	4.38
Control	2.38	Case	4.29
Control	2.43	Case	4.05
Control	2.38	Case	4.24
Control	2.52	Case	4.38
Control	2.57		
Control	4.56		
Control	4.30		

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \neq \mu_2 \end{cases}$$

MannWhitney test

Analyze → Nonparametric Tests → Legacy Dialogs → Two Independent Samples test



Mann-Whitney Test

Test Statistics[a]

	Satisfaction
Mann-Whitney U	18.000
Wilcoxon W	154.000
Z	-3.780
Asymp. Sig. (2-tailed)	.000
Exact Sig. [2*(1-tailed Sig.)]	.000[b]

One-Way ANOVA (Analysis of Variance)

- We perform this test when we want to compare the mean of more than two independent groups.
- Assumptions
  1. Independent observations
  2. Normality Normality (Kolmogorov–Smirnov test/ Shapiro–Wilk test)
  3. Homogeneity of variance (Levene's test)

Null hypothesis: All population means are equal.

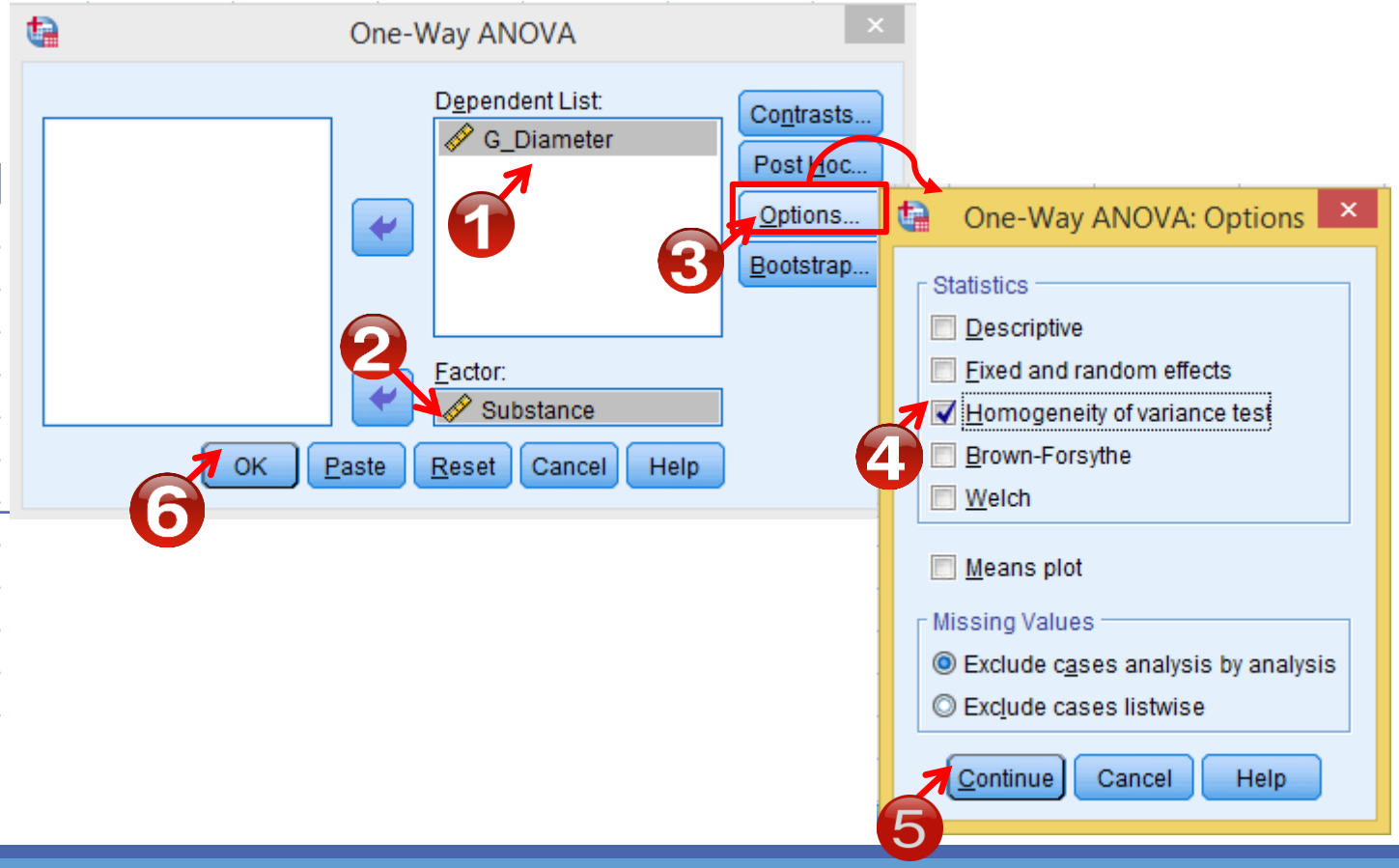
If the p-value is less than alpha, you reject the null hypothesis.

One-Way ANOVA (Analysis of Variance)

Analyze ⟶ Compare Means ⟶ One-Way ANOVA

- Investigating the Factor of Biological Materials on Cell Diameter Size

Substance	G_Diameter	Substance	G_Diameter	Substance	G_Diameter
Esterogen	43.00	Progestron	35.00	Control	20.50
Esterogen	45.00	Progestron	33.50	Control	17.50
Esterogen	44.00	Progestron	29.50	Control	19.50
Esterogen	42.00	Progestron	35.00	Control	23.00
Esterogen	44.00	Progestron	41.00	Control	23.00
Esterogen	41.00	Progestron	35.50	Control	25.50
Esterogen	46.00	Progestron	41.00	Control	16.00
Esterogen	43.00	Progestron	40.50	Control	19.50
Esterogen	46.00	Progestron	41.00	Control	26.00
Esterogen	41.00	Progestron	40.50	Control	23.50
Esterogen	46.00	Progestron	38.50	Control	21.50
Esterogen	42.50	Progestron	42.00	Control	21.00



One-Way ANOVA (Analysis of Variance)

Assumptions (Normality)

Substance = Esterogen

One-Sample Kolmogorov-Smirnov Test[a]

		G_Diameter
N		12
Normal Parameters[b,c]	Mean	43.6250
	Std. Deviation	1.84791
Most Extreme Differences	Absolute	.151
	Positive	.132
	Negative	-.151
Test Statistic		.151
Asymp. Sig. (2-tailed)		.200[d,e]

Substance = Progestron

One-Sample Kolmogorov-Smirnov Test[a]

		G_Diameter
N		12
Normal Parameters[b,c]	Mean	37.7500
	Std. Deviation	3.95141
Most Extreme Differences	Absolute	.257
	Positive	.141
	Negative	-.257
Test Statistic		.257
Asymp. Sig. (2-tailed)		.028[d]

Substance = Control

One-Sample Kolmogorov-Smirnov Test[a]

		G_Diameter
N		12
Normal Parameters[b,c]	Mean	21.3750
	Std. Deviation	3.01606
Most Extreme Differences	Absolute	.122
	Positive	.074
	Negative	-.122
Test Statistic		.122
Asymp. Sig. (2-tailed)		.200[d,e]

Assumptions (Homogeneity of variance)

Test of Homogeneity of Variances

G_Diameter

Levene Statistic	df1	df2	Sig.
4.363	2	33	.021

ANOVA

G_Diameter

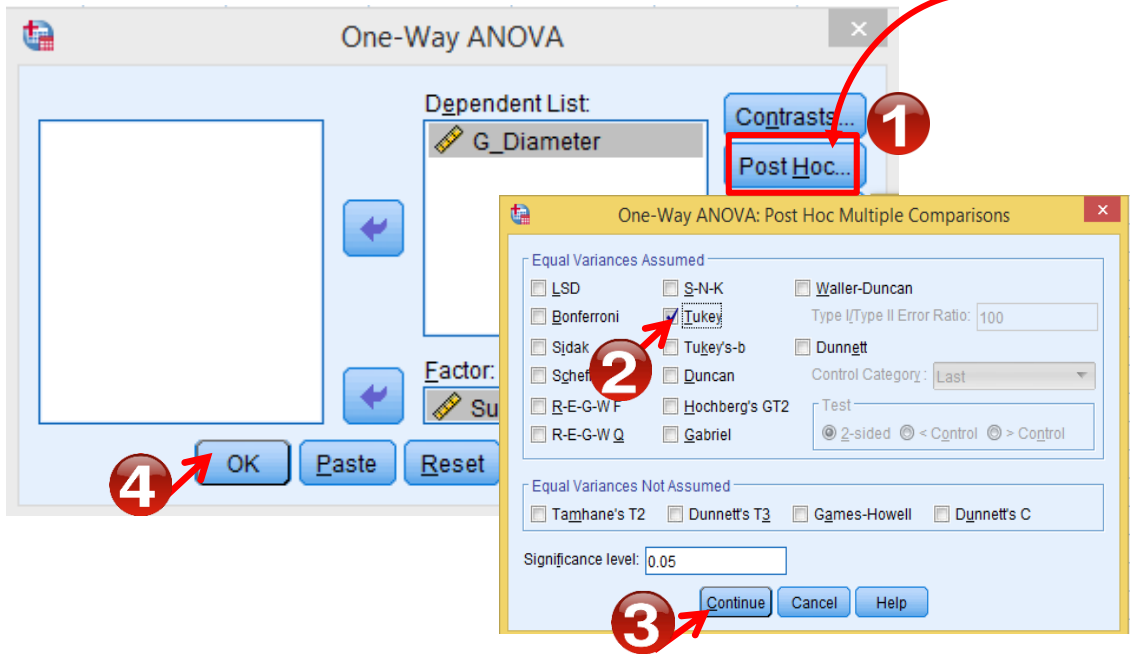
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3190.875	2	1595.438	170.180	.000
Within Groups	309.375	33	9.375		
Total	3500.250	35			

One-Way ANOVA (Analysis of Variance)

ANOVA

G_Diameter

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3190.875	2	1595.438	170.180	.000
Within Groups	309.375	33	9.375		
Total	3500.250	35			



Post Hoc Tests

Multiple Comparisons

Dependent Variable: G_Diameter

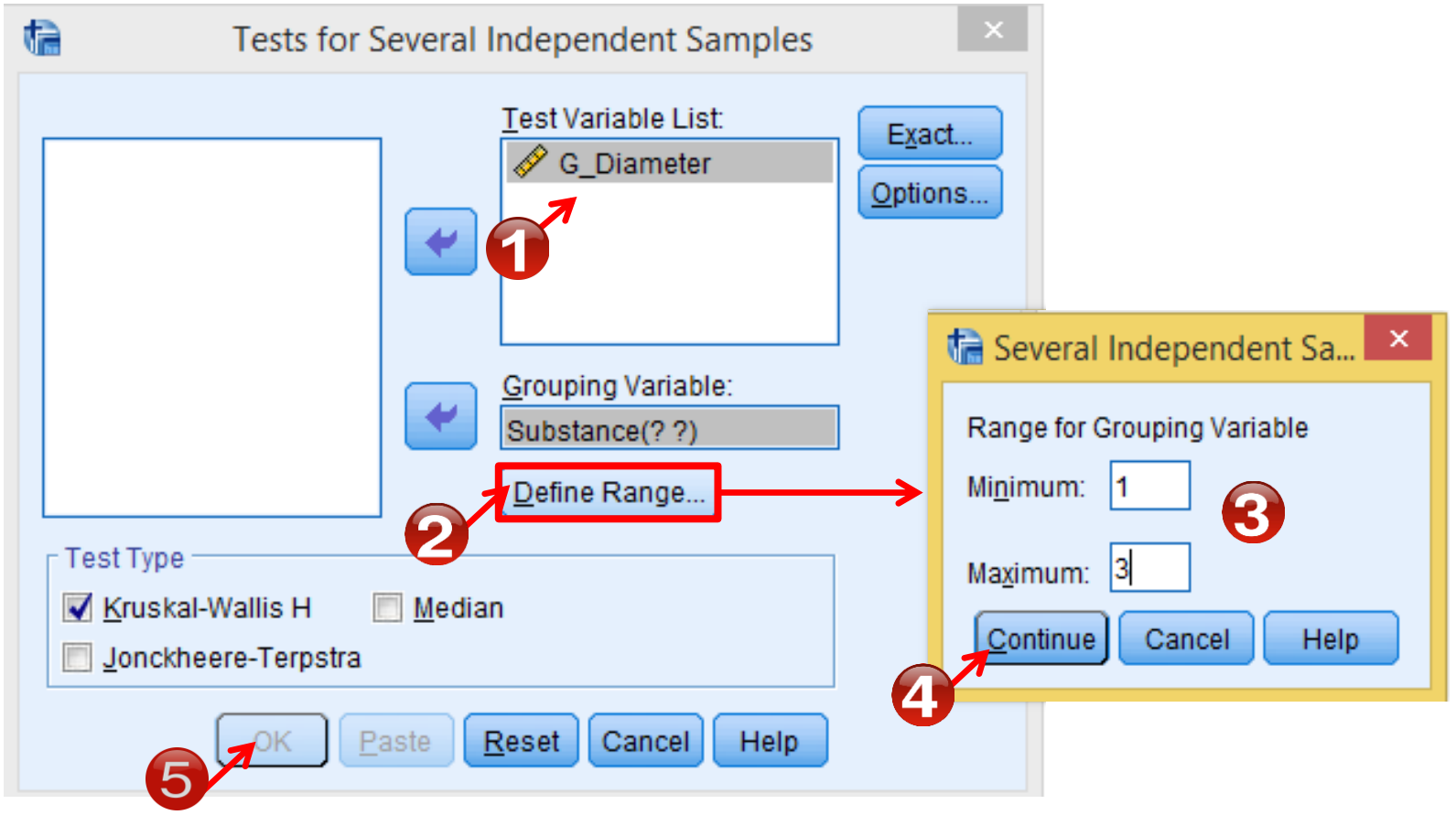
Tukey HSD

(I) Substance	(J) Substance	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Esterogen	Progestron	5.87500*	1.25000	.000	2.8078	8.9422
	Control	22.25000*	1.25000	.000	19.1828	25.3172
Progestron	Esterogen	-5.87500*	1.25000	.000	-8.9422	-2.8078
	Control	16.37500*	1.25000	.000	13.3078	19.4422
Control	Esterogen	-22.25000*	1.25000	.000	-25.3172	-19.1828
	Progestron	-16.37500*	1.25000	.000	-19.4422	-13.3078

*. The mean difference is significant at the 0.05 level.

Kruskal-Wallis Test

Analyze → Nonparametric Tests → Legacy Dialogs → K Independent Samples test



Kruskal-Wallis Test

Test Statistics[a,b]

	G_Diameter
Chi-Square	28.633
df	2
Asymp. Sig.	.000

a. Kruskal Wallis Test

b. Grouping Variable: Substance

Paired sample t-test(Dependent T-Test)

- We perform this test when we want to compares two means that are from the same individual, object, or related units.
- Assumptions
  1. Independent observations
  2. Normality (Kolmogorov–Smirnov test/ Shapiro–Wilk test)

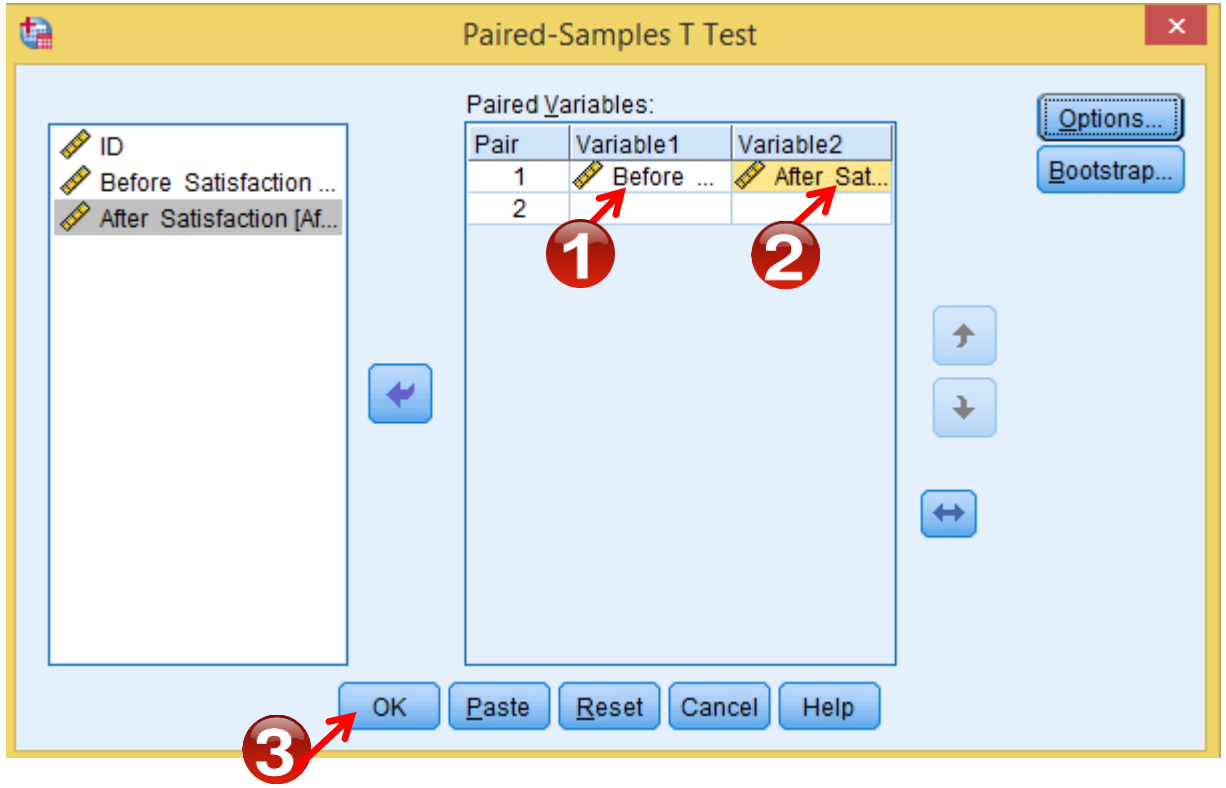
Null hypothesis: The paired population means are equal.
The difference between the paired population means is equal to 0.

If the p-value is less than alpha, you reject the null hypothesis.

Paired sample T-Test(Dependent T-Test)

Analyze → Compare Means → Paired Samples T-Test

ID	Before	After
1	2.29	4.33
2	2.43	4.05
3	2.52	4.19
4	2.52	4.14
5	2.33	4.43
6	2.62	4.38
7	2.33	4.57
8	2.48	4.38
9	2.67	4.38
10	2.38	4.29
11	2.43	4.05
12	2.38	4.24
13	2.52	4.38
14	2.57	4.29
15	4.56	4.05
16	4.30	4.24



Paired sample t-test(Dependent T-Test)

- **Assumptions** (Normality)

One-Sample Kolmogorov-Smirnov Test

		Before Satisfaction	After Satisfaction
N		16	16
Normal Parameters[a,b]	Mean	2.7085	4.2738
	Std. Deviation	.68202	.15109
Most Extreme Differences	Absolute	.399	.136
	Positive	.399	.120
	Negative	-.268	-.136
Test Statistic		.399	.136
Asymp. Sig. (2-tailed)		.000[c]	.200[c,d]

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Before Satisfaction	2.7085	16	.68202	.17051
	After Satisfaction	4.2738	16	.15109	.03777

Paired Samples Correlations

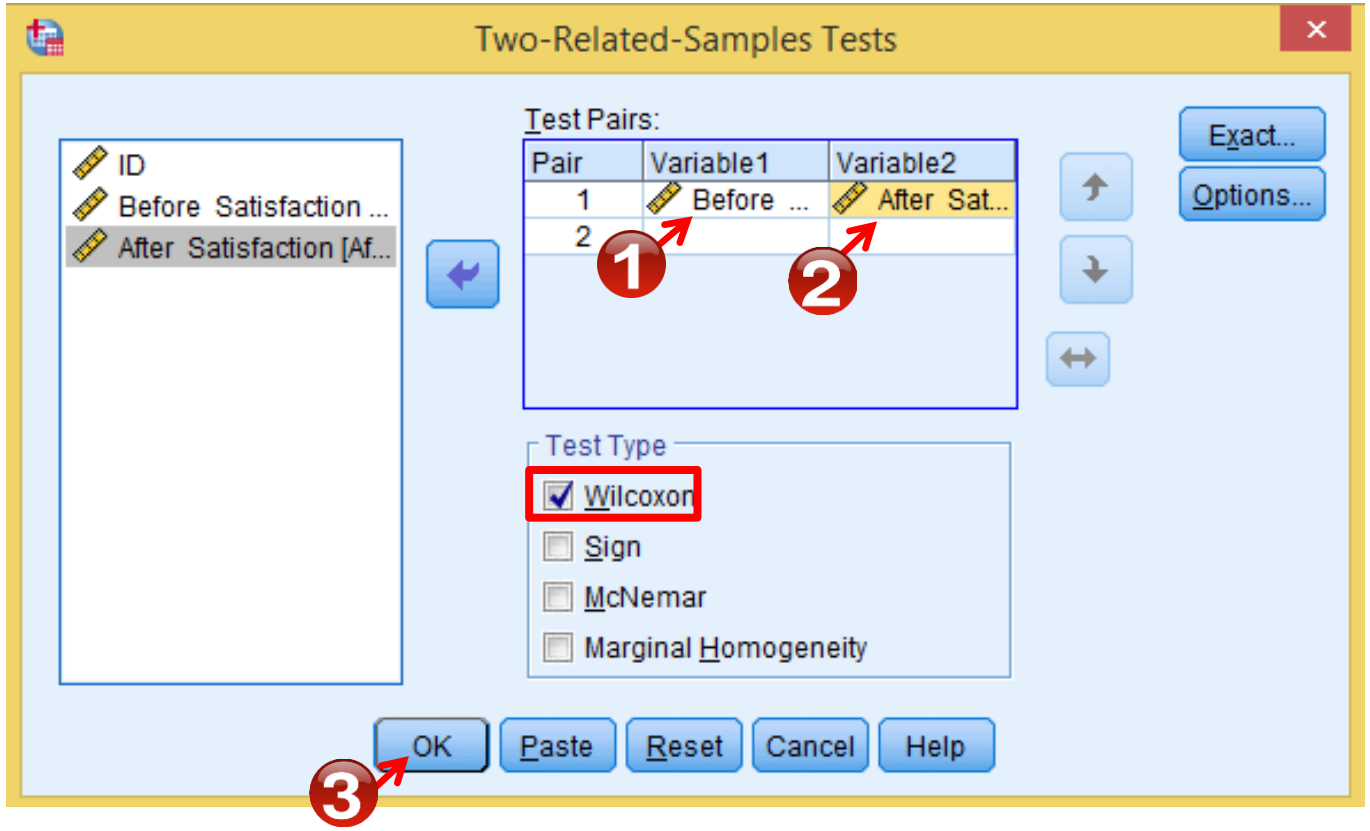
		N	Correlation	Sig.
Pair 1	Before Satisfaction & After Satisfaction	16	-.360	.170

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Before Satisfaction - After Satisfaction	-1.56530	.74982	.18746	-1.96485	-1.16575	-8.350	15	.000

Wilcoxon test

Analyze → Nonparametric Tests → Legacy Dialogs → Two Related Samples test



Test Statistics[a]

	After Satisfaction - Before Satisfaction
Z	-3.365[b]
Asymp. Sig. (2-tailed)	.001

a. Wilcoxon Signed Ranks Test

b. Based on negative ranks.

Statistical Methods of Analysis

Variable	Statistics	Comparison	Test
Nominal	Proportion (P)	2 groups	Fisher's exact test, Chi^2 test
	Proportion (P)	>2 groups	Chi^2 test
Ordinal	Proportion (P)	2 groups	Mann-Whitney
	Proportion (P)	>2 groups	Kruskal-Waillis test
	Proportion (P)	2 groups (paired)	Wilcoxon
Numerical	Mean (μ)	2 groups	*t*-test
	Mean (μ)	>2 groups	ANOVA
	Mean (μ)	2 groups (paired)	Paired *t*-test


Thank
You